DOCUMENT RESUME

ED 349 316 TM 018 609

AUTHOR Marsh, Herbert W.; Balla, John
TITLE Goodness of Fit Confirmatory Factor Analysis: The Effects of Sample Size and Model Parsimony.
PUB DATE 2 Mar 92
NOTE 39p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Computer Simulation; *Estimation (Mathematics); Foreign Countries; *Goodness of Fit; *Mathematical Models; *Sample Size
IDENTIFIERS *Confirmatory Factor Analysis; *Parsimony (Statistics)

ABSTRACT

The influence of sample size (N) and model parsimony on a set of 22 goodness of fit indices was investigated, including those typically used in confirmatory factor analysis and some recently developed indices. For sample data simulated from 2 known population data structures, values for 6 of 22 fit indices were reasonably independent of N and were not significantly affected by the addition of freely estimated parameters known to have 0 values in the population. Indices recommended for evaluating goodness of fit were: (1) two indices based on non-centrality described by R. P. McDonald (1990); (2) a new relative (incremental) index based on non-centrality (P. M. Bentler, 1990 and R. P. McDonald and H. W. Marsh 1990); (3) unbiased estimates of goodness of fit (K. G. Joerskog and D. Sorbom, 1988) presented by J. H. Steiger that are based on non-centrality; and (4) the widely known relative index developed by L. R. Tucker and C. Lewis (1973). Penalties for model complexity designed to control sampling fluctuations and to facilitate the inevitable compromise between goodness of fit and model parsimony were evaluated. Seven tables present analysis results and there is a 36-item list of references. Appendix 1 describes the indices used in the study, and Appendix 2 gives mean values for each model. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Goodness of Fit in Confirmatory Factor Analysis: The Effects of Sample Size and Model Parsimony

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
HERBERT W. MARSH
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 349816

December 17, 1991

Herbert W. Marsh, University of Western Sydney, Macarthur, Australia

John Balla, University of Sydney, Australia

Running Head: Goodness of Fit (printed March 2, 1992)

The author would like to thank Roderick McDonald, David Grayson, John Balla, James Steiger and Stanley Mulaik for helpful comments of on earlier drafts of this paper. The opinions expressed here, however, are those of the author and may not represent those of the people I have acknowledged.

TM018609

BEST COPY AVAILABLE

ABSTRACT

The purpose of the present investigation is to examine the influence of sample size (N) and model parsimony on a set of 22 goodness-of-fit indices including those typically used in confirmatory factor analysis and some recently developed indices. For sample data simulated from two known population data structures, values for 6 of 22 fit indices were reasonably independent of N and were not significantly affected by the addition of freely estimated parameters known to have zero values in the population: two indices based on noncentrality described by McDonald (1990; McDonald and Marsh, 1990), a new relative (incremental) index based on noncentrality (Bentler, 1990; McDonald & Marsh, 1990), unbiased estimates of LISREL's GFI and AGFI (Joreskog & Sorbom, 1988) presented by Steiger (1989, 1990) that are based on noncentrality, and the widely known relative index developed by Tucker and Lewis (1973). Penalties for model complexity designed to control for sampling fluctuations and to facilitate the inevitable compromise between goodness of fit and model parsimony were evaluated.

An important, unresolved issue in structural equation modeling is how to evaluate a model and how to select among competing models. On the basis of theory, previous research, and knowledge of the data, the researcher typically posits a set of alternative models to explain relations among the measured variables. Model selection and evaluation is based on a subjective combination of substantive issues, inspection of parameter estimates, goodness of fit, parsimony, interpretability, and a comparison of the performances of competing models. Whereas evaluating goodness of fit is the focus of this investigation, it is important to emphasize that this is only one aspect of model evaluation. For example, if multiple indicators of the same factor load in different directions when they are predicted to load in the same direction, then the results do not support the construct validity of the measures even if the goodness of fit is otherwise good. Tests of statistical significance and fit indices are useful for model evaluation, but ultimately researchers must also rely on nonstatistical criteria to resolve the inevitable compromise between model parsimony and goodness of fit. As noted by MacCallum (1990, p. 161): "In the end, of course, one must not rely solely on indicators of model fit in evaluating the appropriateness of a model. Support of a model must also be based on its theoretical meaningfulness, interpretability, and parsimony."

Issues and Potential Problems in Assessing Fit

For present purposes it is relevant to distinguish between the true population covariance matrix (Σ), the sample covariance matrix (S), the fitted covariance matrix (Σ_E) derived from Σ for the "best fitting" parameters based on the model that is being tested, and the fitted covariance matrix (E) derived from S. Thus, S is an unbiased estimate of Σ and E is an unbiased estimate of Σ_E. In practice the problem of goodness of fit is how to decide whether E and S are sufficiently similar to justify the conclusion that the specific model used to generate E adequately fits a particular set of data. I refer to the difference between E and S as the empirical discrepancy. Following Cudeck (1991), it is useful to distinguish between (a) estimation discrepancy based on differences between E and Σ_E that reflect sampling fluctuations and (b) approximation discrepancy based on differences between Σ and Σ_E. As sample size increases, S approaches Σ so that estimation discrepancy approaches zero. Approximation discrepancy, however, only involves Σ and Σ_E so that its value does not vary with N. This distinction is important because some indices considered here estimate disprepancy due to approximation that is independent of N, whereas others reflect the combined influences of approximation discrepancy and estimation discrepancy that varies systematically with N.

The classical form of statistical hypothesis testing is generally inappropriate for evaluating the size of the empirical discrepancy as noted by many researchers (e.g., Bentler & Bonett, 1980; Cudeck, 1991; Cudeck & Browne, 1983; Joreskog & Sorbom, 1988; Lawley & Maxwell, 1971; MacCallum, 1990; Marsh, Balla & McDonald, 1988; McDonald, 1985; McDonald & Marsh, 1990; Steiger & Lind, 1980). The failure to obtain a nonsignificant X^2 may reflect a poorly specified model, the power of the test, or the failure to satisfy other assumptions underlying the statistical test. Accordingly, a large number of fit indices that are a function of the

empirical discrepancy have been proposed (e.g., Akaike, 1974; 1987; Bollen, 1986; Bentler & Bonett, 1980; Bozdogan, 1987; Cudeck, 1991; Cudeck & Browne, 1983; Hoelter, 1983; James, Mulaik & Brett, 1982; Joreskog & Sorbom, 1988; Marsh, Balla & McDonald, 1988; McArdle, 1986; McDonald, 1990; McDonald & Marsh, 1990; Mulaik, et al. 1989; Schwartz, 1978; Steiger & Lind, 1980; Steiger, 1989, 1990; Tanaka, 1987; Tanaka & Huba, 1986; Tucker & Lewis, 1973) for evaluating fit and comparing alternative models. The purpose of the present investigation is to examine how goodness of fit as assessed with a variety of indices is related to N, the number of cases in the data to be fit, and model parsimony.

Comparative Fit

Bentler (1990; Bentler & Bonett, 1980) noted the usefulness of testing a series of nested models M_0,..., M_i,..., M_j, ..., M_k, ..., M_s in which M_0 is a suitably defined null model, M_s is the saturated model with df=0 and M_i, M_j, and M_k are models with positive df of intermediate complexity. Any two models are nested so long as the set of parameters estimated in the more restrictive model is a subset of the parameters estimated in the less restrictive model. Under appropriate assumptions, the difference in X^2s between two nested models has a X^2 distribution and so can be tested in relation to statistical significance. This significance test of the difference between two nested modesl has essentially the same strengths and weaknesses of the X^2 test applied to any one model. For this reason, Bentler (1990) recommends that goodness of fit is scaled along a 0-to-1 continuum in which the endpoints are defined by M_0 and M_s respectively. This provides a potentially useful frame of reference against which to evaluate the fit of any one model or the difference between a pair of models. Because of the appeal of relative fit indices like those described by Bentler (1990), Marsh, Balla, and McDonald (1988), and McDonald and Marsh (1990) it is useful to evaluate further their strengths and weaknesses.

Systematic Bias Related to N.

I assert that one useful characteristic for a fit index of emprical discrepancies between S and E based on sample data is to provide an unbiased estimate of the corresponding approximation discrepancies between Σ and Σ_E in the population. If the expected value of a sample statistic varies systematically with N, then the statistic provides a biased estimate of the corresponding population parameter. Bias in a fit statistic is undesirable because, on average, tests of a given model for given variables might yield radically different conclusions at different sample sizes and because, on average, decisions based on sample data may differ systematically from those based on the population. This influence of N can be demonstrated empirically as here or mathematically as in McDonald and Marsh (1990). McDonald and Marsh (1990) derived the expected values of most of the fit statistics considered here, demonstrating that many indices based on finite N are biased estimates of their asymptotic limits and that the size of this bias is a systematically decreasing function of N^{-1}. For different indices the direction of this bias was sometimes positive and sometimes negative. The present investigation uses a Monte Carlo approach to evaluate the extent of this bias.

It is important to distinguish between the effect of N on the value of fit indices that is the present focus and the effect of N on the power of tests used to evaluate fit or specific parameters within a model (e.g., Kaplan, 1990; Satorra, 1989; Satorra & Saris, 1985) and specification error search strategies used to modify misfit models (e.g., Kaplan, 1990). Moreover, a "good" estimator may have desirable properties other than being unbiased. In some statistical applications, it may be preferable to use a biased estimator, if the extent of bias is trivial and the sampling variability of the biased statistic is much smaller than that of the corresponding unbiased statistic (see Bentler, 1990; Steiger, 1990). Whereas these important issues warrant further consideration, they are beyond the already broad scope of the present investigation.

Penalites for Model Complexity: Parsimony and Estimation Discrepancy.

The inclusion of additional parameters, particularly when based on a posteriori criteria and tested with the same data, may provide an "illusory" improvement in fit. Researchers have approached this problem from different perspectives. For example, Steiger (1989; Steiger & Lind, 1980) noted that indices that fail to compensate for model complexity inevitably lead to the selection of the most complex model even when simpler models fit the data nearly as well. James, et al., (1982, p. 155) ask "how efficient is the increase in fit going from the null model with many degrees of freedom to another model with just a few degrees of freedom in terms of degrees of freedom lost in estimating more parameters?" Joreskog and Sorbom (1981, p. I. 40) note that when the change in X^2 is close to the difference in df due to the addition of new parameters, then the "improvement in fit is obtained by 'capitalizing on chance,' and that the added parameters may not have real significance and meaning." Cudeck and Browne (1983; also see Marsh, 1987) described the method of cross-validation to determine the ability of a set of parameter estimates to adequately describe data based on new observations from the same population. Cudeck and Browne also used variants of the AIC (Akaike, 1974) and SC (Schwartz, 1978) indices for this purpose and recently derived a single-sample cross validation index (Browne and Cudeck, 1989). Bozdogan (1987) noted that model selection requires researchers to achieve an appropriate balance between problems associated with "overfitting" and "underfitting" the data, and that different fit indices vary in the balance of protection that they offer from these competing risks. McDonald and Marsh (1990) noted the need to strike a balance between badness of fit and model complexity or, equivalently, between goodness of fit and model parsimony.

Following James, et al. (1982) and Mulaik et al. (1989), I operationalize parsimony as the ratio of degrees of freedom in the model being tested and degrees of freedom in the null model (e.g., the number of covariances below the main diagonal in the variance/covariance matrix). Conversely, I operationalize model complexity to be 1 minus the parsimony ratio.[2]

Penalties for model complexity have different implications for fitting a known population covariance matrix Σ and a sample covariance matrix S. Even if Σ is known, freeing enough parameters still leads to a perfect fit. As noted by McDonald and Marsh (1990) the compromise between goodness of fit (approximation

discrepancy) and parsimony is not an issue of sampling in that even if Σ is known, an appropriate compromise is still required. Consistent with this perspective, I reserve the term parsimony penalty to refer to penalties for model complexity that are applied to indices based on unbiased estimates of approximation discreancy (between Σ and Σ_E) and do not vary with N. Although there are many ways to operationalize parsimony penalties, one popular approach is embodied in the the parsimony indices described by James, et al. (1982) and Mulaik et al. (1989). Because of the intuitive appeal of this approach and the attention that it has attracted, it is useful to evaluate further its strengths and weaknesses.

When population parameters are estimated from a finite S, parameter estimates will fit the data used to estimate the parameters better than they will fit the data derived from a cross-validation sample or from the population from which the sample came. For example, even if a parameter is known to have a zero value for Σ, its estimate generally will not have a zero value for a finite S. Hence, estimating this parameter value leads to an improved fit in S but not in Σ. This tendency is greatly magnified when modifications to the a priori model are based on the same data used to test both the a priori and a posteriori models, and when N is small. This is an issue of sampling fluctuations because estimation disprecancy is zero when when a model is fit to a true population covariance matrix and tends to zero as N increases sufficiently for estimates based on sample data. For present purposes I use the term estimation penality to refer to penalities for model complexity that reflect estimation discrepancy (between E and Σ_E) and that vary systematically with N. The distinction between parsimony penalties and estimation penalties is important because some indices are aimed primarily at controlling for over-fitting in sample data whereas others are intended to facilitate the compromise between model parsimony and approximation discrepancy.

Estimation penalties are incorporated into a set of information criterion indices described by Browne and Cudeck (1989), Cudeck and Browne (1983), and Cudeck and Henly (1991). These indices integregate information about approximation discrepancy and estimation discrepancy into a single index. The information criterion indices lead to the selection of less complex models for sufficiently small Ns, increasingly more complex models for larger Ns, and the saturated model for a sufficiently large N. Support for these indices comes from the theoretical application of information theory (e.g., Akaike, 1974; Bozdogan, 1987) and empirical demonstrations that models selected by these indices tend to cross-validate better than alternative models (Cudeck & Browne, 1983; Cudeck and Henly, 1991).

Superficially, the rationale underying the information criterion indices appears to conflict with the assertion that an index of empirical discrepancy should provide an unbiased estimate of approximation discrepancy. This led Cudeck and Henly (1991) to counter the assertion by McDonald and Marsh (1990) and Marsh, Balla, and McDonald (1990) that a systematic relation between the value of an index and the sample size on which it was based was an undesirable feature in estimates of approximation discrepancy. It should be emphasized, however, that the intent of the information criterion indices is not to estimate approximation

discrepancy, but to penalize approximation discrepancy appropriately for estimation discrepancy. Thus the disagreement is not in the contention that estimations of approximation discrepancy based on samples should provide unbiased estimates of the corresponding population values, but what the appropriate role of estimation discrepancy should be in evaluating approximation discrepancy. Hence, the role of estimation discrepancy in evaluating and selecting models will be considered here in an attempt to clarify this apparent controversy.

Overview of the Present Investigation

The issues related to goodness of fit and penalties for model complexity are blurred in practice because Σ is not known and it is difficult to operationalize terms such as over- or under-fitting. These terms can easily be operationalized, however, for simulated data derived from a known Σ in which some of the population parameters are defined to be zero. If a model is fit to Σ, allowing these parameters to differ from zero will not improve the fit. When the model is fit to a finite S, however, allowing these parameters to differ from zero will improve the fit. For purposes of this simulation study, over- and under-fitting are operationalized as freely estimating parameters known to be zero in the population and fixing at zero parameter estimates known to be non-zero in the population. In this special case, the effects of over- and under-fitting can be evaluated in relation to alternative fit indices.

Our objective is to examine the effects of model parsimony and of N on a set of 22 fit indices (Appendix 1). Data were generated for two different population data structures and a variety of models used to fit the data were developed in relation to the known population parameters. Some models posited parameters to be zero that were known to be non-zero in the population, thus providing models that were under-fit. Other models freely estimated values for parameters that were known to be zero in the population, thus providing models that were over-fit. Covariance matrices to be fit by the alternative models were based on one of six different sample sizes varying from 50 to 1600.

For present purposes the set of 22 goodness of fit indices (Appendix 1) are classified into three types, namely: (a) stand-alone (absolute) indices, (b) incremental (relative) indices, and (c) parsimony indices. The 16 stand-alone indices are based on the results of just a target model, the a priori model posited by the researcher to fit the data. These indices are provided by, or easily computed from results provided by, LISREL and most other statistical packages used to fit structural equation models. The incremental indices are based on the difference between the target model and an alternative model such as a "null" model in which Σ is posited to be a diagonal matrix (Bentler, 1990; Bentler & Bonett, 1980; Marsh, Balla & McDonald, 1988). A parsimony index corresponding to any incremental index can be obtained by multiplying the parsimony ratio times the value of the index, and Mulaik et al. (1989) specifically note the advantages of using an unbiased incremental index to define parsimony indices. For present purposes, two parsimony indices are evaluated; one based on the Bentler-Bonett Index (BBI; Bentler & Bonett, 1980) that has been used widely for this purpose and a new index derived from the relative noncentrality index (RNI).

Method

The CFA Model and Analyses

All analyses were conducted with LISREL V (Joreskog & Sorbom, 1981) using the method of maximum likelihood. In each of the analyses involving 9 observed variables (Table 1) a set of eight substantive models (M1 - M8; see Table 2) posited between 18 and 33 parameters (K) to define 1, 2, or 3 factors. Hence the df (.5 x 9 x 10 - K) varied from 27 to 12. M1 - M8 and their relation to the population parameters used to generate the data (Table 1) are summarized in Table 2. A null model was also tested for each covariance matrix positing a diagonal Σ that had 36 (.5 (10 x 9) - 9) df. The eight substantive models and the null model, were tested for each of 120 covariance matrices described below.

Insert Tables 1 & 2 About Here

The Data.

The Sample Sizes. The six sample sizes to be considered (50, 100, 200, 400, 800 and 1600) were selected to span the range of sample sizes typically encountered in CFA. For each of the two population data structures to be considered, ten random samples were generated for each sample size and the same nine models were fit to these 120 (2 population data structures x 6 sample sizes x 10 cases) covariance matrices.

Simple structure simulated data (SSIM). The nine measured variables were defined with the random number generator from the commercially available SPSS package (Hull & Nie, 1981) based on population values presented in Table 1. Each variable was defined to reflect only one factor (factor loadings were .6, .7 or .8) and a normally distributed random error component, and the three factors were defined to be correlated. A total of 31,500 cases were generated and divided into 60 sets of data such that each sample size was represented by 10 covariance matrices. M1-M8 and the null model were fit to each of the 60 covariance matrices.

M8 (see Table 2) freely estimates all parameters known to be nonzero for the SSIM population data structure and fixes at zero all parameters known to be zero for the SSIM population data structure. Hence this model is a "true" model in that it is able to reproduce perfectly the population parameters used to generate the data and it is also the most restrictive (i.e., parsimonious) true model considered. For present purposes I refer to M8 as the "correct" model for the SSIM data. Models positing only one or two factors (M1 and M2 in Table 2) are incorrect (i.e., have nonzero approximation discrepancy) and this should be reflected in the values of the fit indices. M8 is nested in each of the remaining five substantive models in that M3-M7 contain all the nonzero parameters in M8 along with a varying number of additional parameters. These additional parameters are superfluous in that their population values, the values from the population data structures used to generate the data, are zero and are fixed to be zero in M8. The fit indices of these over-fit models are used to evaluate how various indices are affected by overfitting. To the extent that any of these over-fit models fit the SSIM data significantly better than M8 according to a particular index, then the index does not control adequately over-fitting. To the extent that any of these models fit the data significantly poorer than M8 according to any

particular indices, then, perhaps, the index can be said to over-compensates for the inclusion of additional freely estimated parameters known to be zero in the population. This relation between the substantive models to be tested and the SSIM data was the basis of a priori contrasts used to compare various models (see Table 3).

Insert Table 3 About Here

Complex Structure Simulated Data (CSIM). The nine measured variables were defined as with the SSIM data in that each was substantially related to one of three factors. Additionally, however, six of the nine measured variables -- two for each factor -- were weakly related to one other factor. Hence, the population data structure had 15 nonzero factor loadings: 9 "major" loadings varying between .6 and .8 as in the SSIM data and 6 "minor" factor loadings all equal to .2 (see Table 1). The value of .2 was selected so as to be small, but to be large enough to be substantively important in many applications. Again, a total of 31,500 cases were generated and divided into 60 sets of data such that each sample size was represented by 10 covariance matrices, and the null and hypothesized models were fit to these 60 covariance matrices.

M7 (see Table 2) freely estimates all parameters known to be nonzero for the CSIM population data structure and fixes at zero all parameters known to be zero for the CSIM population data structure. Thus M7 is the most restrictive true model considered here and I refer to M7 as the "correct" model for the CSIM data. M3, positing three unrestricted factors, should also be able to fit the data though it is less parsimonious. Models positing only one or two factors (M1 and M2 in Table 2) should not be able to fit the data. In each of the remaining models positing three factors, either 3 minor factor loadings (M4 and M5) or all 6 minor factor loadings (M6 and M8) are constrained to be zero. Of these four models, only M6 freely estimates parameters whose population values are known to be zero. This set of models provides additional tests of models known to over-fit or under-fit the data in relation to the known population parameters. Two sets of models (M3 and M7, and M6 and M8) are equivalent in their ability to fit the data (i.e., have equal approximation discrepancy) but differ in the number of parameters that are estimated. For two additional sets of models (M3 vs. M8; M4 and M5 vs. M8), the model that should fit best requires more parameters, thus providing a test for over-correcting for the inclusion of freely estimated parameters known to be zero in the population. This relation between the substantive models to be tested and the CSIM populations data structure was the basis of a priori contrasts used to compare various models (see Table 3).

Analyses. The analyses are based on differences in fit indices associated with the set of 8 (substantive models) x 6 (sample sizes). Repeated measures ANOVAs were conducted in which the 6 sample sizes were levels of a between subjects factor and, because each model was estimated from the same data sets, the 8 substantive models were levels of the repeated measures variable. Tests of statistical significance of the repeated measures variables used the conservative Greenhouse-Geisser (SPSS, 1990) correction. The major emphasis, however, was on estimates of variance explained that do not depend on assumptions underlying the

BEST COPY AVAILABLE

tests of statistical significance. These overall analyses were followed up by the set of 9 a priori contrasts involving comparisons of different models, each based on a single df, that are described in Table 3. Because each contrast was based on a single df, the assumption of sphericity in this repeated measures variable is automatically met. Separate analyses were conducted for each of the 22 fit indices and separate analyses were conducted for results of the SSIM and CSIM data. Mean values for each cell of this design are presented in Appendix 2 for all 22 fit indices.

Results

Simple Simulated (SSIM) Data.

Sample Size (N) Effect. The effect of N is statistically significant and substantial (eta > .1) for 15 of the 22 indices and can be explained primarily by the linear effect of log N (Table 4). The relation between goodness of fit and N is nonsignificant for five indices based on noncentrality (Dk, Mc, GFI*, AGFI*, RNI), and small -- though significant -- for the noncentrality parsimony index (PRNI) and the TLI incremental index. An inspection of the sample means (Appendix 2A) further clarifies the nature of the effect of N.

Models. In relation to the known population parameters for the SSIM data, M3-M8 should be able to fit the data (i.e., all nonzero population parameters are freely estimated) whereas M1 and M2 should not. For all 22 fit indices there are significant differences in the ability of competing models to fit the data (see Etas attributable to the Model in Table 4), and most of this difference is due to the poorer fits of M1 and M2.

For the SSIM data, M3 - M8 are all able to fit the data but differ in the number of parameters known to be zero in the population that are estimated freely. For analysis conducted on just M3 - M8 (Table 5) 8 indices show significantly better fits when more parameters known to be zero are estimated, 5 indices show significantly poorer fits when more parameters known to be zero are estimated, and the remaining 9 indices are not significantly related to the number of parameters known to be zero that are freely estimated.

Insert Table 4 About Here

The purpose of the a priori contrasts (Table 4) is to test the ability of the 22 indices to differentiate among models known to differ in their ability to fit the data. For the SSIM data, the set of 9 a priori contrasts can be divided into two types. Contrasts 1 and 2 compare models that are known to differ substantially in their ability to fit the population data, whereas contrasts 3 - 9 compare models that are all able to fit the population data.

Contrasts 1 and 2 should be statistically significant and favor the models better able to fit the population data.. For 20 of the 22 indices, contrasts 1 and 2 are statistically significant and in the right direction. For the two parsimony indices one of the contrasts is significant but in the wrong direction. Inspection of Appendix 2A shows that at all sample sizes the parsimony indices lead on average to the selection of a more parsimonious model that is not able to fit the data (M2) over a less parsimonious but true model (M3). These

results based on contrasts 1 and 2 provide support for 20 of the indices, but may call into question the usefulness the parsimony indices.

The SSIM data was generated by a population model containing nonzero values for only the 21 parameters estimated in M8 (Table 2) and all other parameters are known to be zero. Contrasts 3, 4, 6, 7, 8 and 9 all compare models that are able to fit the population data but differ in the number of parameters known to be zero that are freely estimated. For each of these contrasts (Table 4), a plus (+) indicates that the model with more parameters known to be zero fits the data better whereas a minus (-) indicates the opposite. The behavior of the different fit indices in relation to these contrasts vary substantially and fall into three classifications.

1) For nine fit indices (X^2/df, RMS, AGFI, GFI*, AGFI, DK, MC, TLI, and RNI), these contrasts are small and typically not statistically significant. That is, the values of these indices do not change substantially for models freely estimating more parameters known to be zero.

2) For eight indices (FF, LHR, X^2, RMR, GFI, FFI1, BBI, and, perhaps, CN) all statistically significant contrasts favor the models that posit more parameters known to be zero. For these indices, over-fit models positing more parameters known to be zero fit the data better on average. Because the true population parameters for these additional parameters are known to be zero for this simulated data, this improved fit is a function of over-fitting sample data.

3) For five fit indices (AIC*, SC*, Ck, PBBI, and PRNI), all statistically significant contrasts favor models that posit fewer parameters. That is, models positing more parameters known to be zero fit the data more poorly than models positing fewer parameters so that these indices can be said to penalize model complexity. The potential danger in penalizing model complexity too severely is observed for the parsimony indices in relation to contrasts 1 and 2; these indices so severely penalized model complexity that a true model had a significantly poorer value than more parsimonious models that were not true. Examination of the contrasts for the remaining three indices suggests that the SC* penalizes model complexity more severely than AIC* and Ck, but there is no basis for claiming that model complexity is penalized too severely.

Complex Simulated (CSIM) Data.

Sample Size (N) Effect. The effect of N is statistically significant and substantial for 14 of the 22 indices and most of this effect is linearly related to log N (see Table 6 and Appendix 2B). Eight indices (RMS, GFI*, AGFI*, Dk, Mc, TLI, RNI, and PRNI) are not significantly related to N. These results are consistent with earlier analyses based on the SSIM data.

Insert Table 6 About Here

Model Effects. For the CSIM data the effect of the different models is statistically significant and substantial for all 22 fit indices (Table 6). The set of 9 a priori contrasts designed to examine these effects can be divided into two types. In relation to approximation discrepancies, contrasts 1, 2, 4, 5, 6, 7, and 9 compare

models known to differ in their ability to fit the population data, whereas contrasts 3 and 8 compare models that are equally able to fit the population data but differ in the number of parameters known to be zero that are freely estimated.

Contrasts 1 and 2 compare the 3 unrestricted models positing 1, 2 and 3 factors. These are gross tests in that comparisons are based on models that differ substantially in their ability to fit the population data. For 20 of the 22 indices, contrasts 1 and 2 are statistically significant and in the right direction. For the two parsimony indices both contrasts are in the wrong direction and one is statistically significant. Hence, these indices penalize model complexity too severely. These results based on contrasts 1 and 2 are similar to findings based on the SSIM data (also see Appendix 2B) and so are not discussed further.

Contrasts 4 and 9 are also rather gross tests in that models that are able to fit the population data (M3 and M7) are compared to M8 in which all 6 minor factor loadings known to be nonzero in the population model are fixed to be zero. For 18 of the 22 indices, these contrasts are statistically significant and in the right direction. For SC* and the two parsimony indices, both of these comparisons is statistically significant and in the wrong direction, whereas one constrast for Ck is not statistically significant. This suggests that with respect to these contrasts, the two parsimony indices, SC*, and, perhaps, Ck may penalize model complexity too severely.

Contrasts 6 and 7 are less gross in that the models being compared differ in terms of only 3 of the 6 minor factor loadings. For 19 of the 22 indices, these contrasts are statistically significant and in the right direction. For the parsimony indices, both these contrasts are statistically significant but in the wrong direction. For SC* both contrasts are in the wrong direction and one is statistically significant. This suggests that with respect to these contrasts, SC* and the two parsimony indices may penalize model complexity too severely.

Contrasts 3 and 8 compare models that are equally able to fit the data but differ in the number of freely estimated parameters known to be zero. The behavior of the indices fell into three general categories; for 9 indices both of these contrasts are statistically significant such that the model positing more parameters fits the data better; for 8 indices neither contrast is statistically significant; for 5 indices one or both of these contrasts are statistically significant such that the model positing fewer parameters fits the data more poorly.

Discussion

Results for both the SSIM and the CSIM population data structures lead to clear conclusions about the behavior of fit indices considered here. Six indices (GFI*, AGFI*, Dk, Mc, TLI, RNI) were not substantially related to N and were not significantly affected by freely estimating additional parameters that had population values known to be zero. Whereas the results are clear, there is room for disagreement as to their interpretation. McDonald and Marsh (1990) discuss some of the relevant theoretical issues, and other perspectives are offered by Mulaik et al. (1989), Bentler (1989), Bozdogan (1987), Cudeck and Henly (1991), Cudeck and Browne

(1983; Browne & Cudeck, 1989), and Steiger (1989, 1990). Two possible criteria of acceptable fit indices are motivating assumptions of the present investigation:

1. The value of an index of empirical discrepancy based on S should be an unbiased estimate of the approximation discrepancy based on Σ. This follows because: (a) when applied to sample data, indices not enjoying this property may lead investigators to select models that differ systematically and are incompatible with decisions based on the population and (b) investigators using different sample sizes may systematically choose different and incompatible models for the same variables.

2. If a restrictive model with positive df is able to describe the population and a series of nested models including the most restrictive true model is fit to sample data, a goodness of fit index should lead to the selection of the most restrictive true model.

It could be claimed that all models containing additional freely estimated parameters known to be zero in the population beyond those of the most restrictive true model are equally capable of fitting the data and should therefore result in equal estimates of fit. This follows in that the approximation discrepancy of all such models are equal. Conversely, it could be claimed that a goodness of fit index that decreases with the addition of parameters known to be zero is penalizing model complexity appropriately. A number of indices considered here are designed to facilitate a compromise between model complexity and goodness of fit through the application of parsimony penalties or estimation penalties. An examination of some of these indices in relation to the present results demonstrates that there are substantial differences in the behavior of these penalty functions.

Penalties For Model Complexity

RMS, AGFI*, and TLI. RMS, AGFI*, and TLI incorporate parsimony penalties. McDonald and Marsh (1990; also see Marsh, Balla & McDonald, 1988) and Steiger (1989, 1990) showed that the RMR, GFI, and AGFI indices provided by LISREL are biased estimates of their asymptotic population values. Steiger (1989, 1990) described alternative indices (RMS, GFI*, and AGFI*; see Appendix 1) that he suggested were not systematically related to N. RMS and AGFI* penalize model complexity, whereas GFI* does not. McDonald and Marsh (1990) demonstrated that the TLI could be expressed in terms of the Dks for a the model to be tested and a suitably defined null model, and the inverse of James al.'s parsimony ratio (see Appendix 1; also see Bentler, 1990). In this respect, TLI can also be said to penalize model complexity. RMS, AGFI*, and TLI are undefined due to division by zero for the saturated model with df=0 (see Appendix 1). Even when df approaches zero, df is a noncontinuous step function that is specific to the particular application and the indices may be unstable (see Bentler, 1990). TLI is also undefined when Dk for the null model is zero and may be unstable when Dk approaches zero.

The AGFI* and TLI were unrelated to N, whereas there was a small relation between N and RMS that requires further consideration. The RMS is based on the square root of Dk. Noting that Dk is in the metric of

squared residuals, Steiger (1989) used the square root of Dk so as "to return the index to the same metric as the original parameters" (p. 81). Because Dk can be negative due to sampling fluctuations, Steiger (1989; see Appendix 1) set negative values of Dk equal to zero for purposes of defining RMS. The effects of this decision rule, however, vary systematically with N since the size of sampling fluctuations vary with N. The negative values of Dk are likely to be more negative on average for small N than for large N so that Steiger's decision rule results in a more substantial truncation of the negative tail of the distribution of Dk values for small N. Consistent with this observation, for all true models (see Table 5 and Appendix 2) RMS is systematically related to N even though Dk is not. Whereas this problem with RMS is obvious for true models used to fit simulated data, it would have a similar, although smaller, effect for any models providing a reasonably accurate representation of real data. It should be noted, however, that the problem with RMS identified here is exaggerated by the use of models that fit the data so well. The size of the bias, although clearly evident here, may be sufficiently small to be unimportant in many applications. Furthermore, this problem may also be clarified by using confidence intervals as proposed by Steiger (1989, 1990). Nevertheless, it may be more useful to simply define RMS in terms of Dk (instead of its square root)[3] or to use some other transformation of Dk (like Mc) that avoids the complication, but these possibilities require further research.

The present results indicated that Dk, RMS, GFI*, AGFI*, and TLI all behaved similarly in that freely estimating additional parameters known to have zero values for Σ did not systematically affect the values of these indices. An apparent paradox is why the penalties for model complexity in RMS, AGFI*, and TLI did not produce results that differed substantially from those based on GFI* and Dk that contained no penalty for model complexity. When Dk = 0 and each of the indices is defined, Dk = RMS = 0 and GFI* = AGFI* = TLI = 1.0. Hence, the penalty functions for model complexity in RMS, AGFI*, and TLI vary inversely with approximation discrepancy and are zero when approximation discrepancy is zero. In this respect, these penalty functions differ from the penalty functions embodied in the parsimony indices and in the information criterion indices that are discussed below. These observations also explain why Dk, RMS, GFI*, AGFI*, and TLI all behave similarly in the present investigation. Most comparisons examined here involved true models (approximation discrepancies = 0) with positive df that varied in the number of freely estimated parameters known to have zero values in the population. For true models, however, RMS, AGFI*, and TLI have no penalty for model complexity so that Dk, RMS, GFI*, AGFI*, and TLI all lead to similar conclusions.

Parsimony indices. The explicit intention of James et al.'s parsimony indices is to combine in one index a measure of goodness of fit and parsimony defined as the ratio of df in the fitted model and an appropriately defined "null" model. For purposes of illustration, consider Models M3, M7, and M8 for the CSIM data (Appendix 2B): M7 is the most parsimonious true model, M3 is a less parsimonious true model, and M8 is more parsimonious than M3 or M7, but is not a true model. The parsimony indices lead on average to the selection of the more parsimonious true model (M7) over the less parsimonious true model (M3) at all N. The two

parsimony indices, however, lead to the selection of the more parsimonious false model (M8) over two less parsimonious true models (M3 and M7). In this special case based on simulated data it can be said that the parsimony indices lead on average to the selection of the wrong model at all sample sizes. For the new PRNI the advantage in favor of the wrong model does not vary with sample size. For the PBBI, however, the advantage in favor of the wrong model is much larger for small N than for large N. This follows in that: (a) the penalty for model complexity -- the parsimony ratio -- does not vary with N; and (b) the asymptotic values of BBI and RNI are the same, but BBI provides a biased estimate of this value that varies with N. Whereas PRNI is apparently more useful than PBBI, these results indicate that parsimony indices may over-penalize model complexity in some situations in relation to criteria established here.

There is no clearly defined minimum parsimony index required to support an adequate fit, but Mulaik (1991) suggested that one would only claim support for models with parsimony indices in the .80's or higher. Any such fixed decision rule, however, creates a number of dilemas. By definition, a parsimony index based on a index of fit that varies between 0 and 1 can never be greater than the parsimony ratio. Consider, for example, the "correct" M8 for the SSIM data. The model fits the data perfectly in the population, has RNI values of approximately 1.0 for sample data, and has a very simple structure (each measured variable loads on one and only one factor) that most would consider to be "parsimonious" (see footnote 2) Yet, the PRNI for M8 (Appendix 2A) is only 0.67. The only way to achieve an acceptable PRNI would be to fix a number of the non-zero parameters values at or near their population values. Thus, for example, if 5 of the 9 non-zero factor loadings were fixed at their known population values (see Table 1) instead of being freely estimated, the PRNI would be .8. Whereas such a practice would be laudable, it is apparently unrealistic for real data and certainly is not common practice.

To further examine this issue of acceptable parsimony indices, I computed the parsimony ratios for a new, hypothetical set of target models differing in the number of factors (1-10) and the number of indicators for each factor (3-8; see Table 7). All models have a very simple (i.e., parsimonious) structure in which each measured variable loads on one and only one factor. As is nearly universal in CFA studies, I assumed that all parameters posited to be nonzero were freely estimated whereas all parameters assumed to be zero are fixed to be zero. Also presented in Table 7 is the minimum RNI needed to achieve a PRNI of .80. Under these circumstances, it is not possible to get a PRNI of .8 for a one factor model based on 8 or fewer indicators even if the fit is perfect (it requires 11 or 12 indicators). For a two-factor model it required 6 indicators per factor and a perfect RNI = 1.0 -- or 7 indicators with RNI = .95 -- to obtain a PRNI of .8. For the three-factor models, it requires at least 5 indicators per factor. For the five- and ten-factor models with many indicators it is interesting to note that the .8 rule of thumb for acceptable parsimony indices is achieved for models that do not meet the RNI > .9 rule of thumb typically applied to relative fit indices. The conventional guidelines for acceptable models based on relative fit and parsimony indices are somewhat arbitrary, based on "current practice" and

implicit, untested assumptions that are difficult to evaluate. Inspection of Table 7 suggests to me, however, that the .8 parsimony index rule makes no sense if an 8-indicator one-factor model that fits the data perfectly is not even close to being "acceptable." Furthermore, the parsimony indices in Table 7 based on one-factor models do not seem comparable with parsimony indices based on five- or ten-factor models. Problems of this sort would also exist for any fixed values of parsimony indices used to define an acceptable fit.

Insert Tables 7 About Here

A second concern in the application of parsimony indices is the trade-off between improved approximation discrepancy and freely estimating additional parameters. For example, with the CSIM data (Appendix 2B) the PNRI led to the selection of M8 (PRNIs about .61) with 24 df over the "correct" M7 (PRNIs about .50) with 18 df (although neither approached the .8 value). Thus, according to the PRNI, the improved goodness of fit achieved by correctly identifying the 6 small nonzero factor loadings of .2 (see Table 1) did not adequately compensate for the penalty due to freely estimating 6 additional parameters. I constructed the data so that the "minor factor loadings" of .2 were small, but large enough to be substantively important in many applications. Whereas there may be contexts in which factor loadings of .2 are not substantively important, as implied by the parsimony indices, the basis of this decision should not be a mechanistic application of parsimony indices. Whereas model parsimony is an important characteristic, further research is needed on how best to operationalize parsimony, on what standards imply acceptable parsimony , and how to integrate integrate information about parsimony, approximation discrepancy, and substantive issues.

AIC*, SC*, and Ck. The development and application of the information criterion indices has taken place primarily within the context of selecting a "best" model from a sequence of nested models in which model complexity varies substantially from some minimal level to that of the saturated model. For example, Browne and Cudeck (1989), Cudeck and Browne (1983) and Cudeck and Henly (1991) have applied these indices to the "number of factors" problem in which a nested sequence of unrestricted factor analysis models positing between 1 and the maximum number of factors (or the saturated model) is evaluated. Other applications include the degree of a polynomial equation needed to fit a function or the order of model complexity in autoregressive time series analyses. In these applications, model selection is a purely empirical attempt to maximize approximation discrepancy subject to the limitations of sampling fluctuations. The intent of the information criterion indices in this situation is to select a "true" model in terms of its order or dimensionality (Bozdogan, 1987). This situation, however, is not analogous to the typical application of covariance structure models in which a limited number of a priori models are proposed that typically have at least moderate parsimony. Hence, there is a need to evaluate the usefulness of information criterion indices in relation to a set of a priori models of intermediate complexity.

AIC*, SC*, and Ck incorporate estimation penalties. The intention of AIC* and SC* is to appropriately balance the risks of underfitting and overfitting a model (e.g., Bozdogan, 1987). Ck, although designed to optimally predict a cross-validation index based on results from a single sample, is closely related to the AIC* (Browne & Cudeck, 1989; Cudeck and Henly, 1991). AIC*, SC* and Ck each have two components; one

BEST COPY AVAILABLE

component reflecting approximation discrepancy that is monotonically related to model complexity and an opposing estimation penalty reflecting estimation discrepancy (sampling error). Dk is an estimate of approximation disrepancy (Cudeck and Henly, 1991) so that the difference between each of these indices and Dk represents the estimation penalty. This comparison is facilitated by expressing each of the information criterion indices as a function of Dk and the corresponding estimation penalty (the expressions in brackets) :

(1) $AIC^* = Dk + [\, df/N + (2K/N) \,]$

$= Dk + [\, (p^* + K)/N \,]$

(2) $SC^* = Dk + [\, df/N + (K \times \ln(N))/N \,]$

$= Dk + [\, (p^* + K (\ln(N) - 1))/N \,]$

(3) $Ck = Dk + [\, df/N + (2K/(N - p - 2)) \,]$

$= Dk + [\, (p^* - K)/N + (2K/(N - p - 2)) \,]$

where K = number of parameters, p = number of variables, $p^* = .5 \times p\,(p+1)$, $df = (.5 \times p\,(p + 1) - K) = p^* - K$.

In this form it is clear that the estimation penalties for all three indices are a monotonically decreasing function of N, are a monotonically increasing function of K, and have asymptotic values of zero. For most applications the penalty associated with SC* will be substantially larger than the other two, whereas the penalty fucntion associated with Ck will be always be somewhat larger than that associated with AIC*. (This behavior is illustrated by the values in Appendix 2.)

Cudeck and Henly (1991) recommended the use of information criterion indices, particularyl Ck, and specifically called into question the claim by Marsh, Balla and McDonald (1988) and McDonald and Marsh (1990) that the influence of sample size is undesirable. The present study, however, further clarifies this apparent controversy. In particular, there seems to be no controversy over the claim that approximation discrepancy should be estimated with an unbiased index such as Dk that is independent of N. Each of the information criterion indices can be expressed as a function of Dk and an estimation penalty based on sampling fluctuations the reflects estimation discrepancy. Hence, the only controversy is the appropriate application of penalties for estimation discrepancy.

Cudeck and Browne (1983), along with many others, noted the limited usefulness of tests of statistical significance for evaluating goodness of fit and indicated the need for alternative indices of fit. The application of information criterion indices, however, is based on a similar logic to that employed in hypothesis testing and often results in the selection of models similar to those based on the X^2 test. For example, McDonald and Marsh (1990) noted that the model selection based on the AIC* in the Cudeck and Browne (1983) study was similar to that which resulted from tests of statistical significance (also see Steiger and Lind, 1980). Inspection of the results reported by Cudeck and Henly (1991), by Browne and Cudeck (1989), and the results presented here (Appendix 2) also demonstrates a similar pattern for model selection based on the Ck index . In order to

demonstrate a mathematical basis for this observation, consider the difference between AIC*s for two nested models where model 1 is more restrictive than model 2. $AIC^*_{diff} = AIC^*_1 - AIC^*_2$ can be written as:

(4) $AIC^*_{diff} = Dk_{diff} + K_{diff}/N$

$= X^2_{diff} + 2\,df_{diff}$

In this form it is clear that the underlying decision rule is to accept the more restrictive model 1 if X^2_{diff} is less than $2\,df_{diff}$ and to accept model 2 otherwise [5]. Comparing these values with those in a standard X^2 table: (a) for $df_{diff} < 7$, the AIC^*_{diff} rule is less conservative than the X^2_{diff} test (i.e., less likely to accept the more restrictive model). (b) for $df_{diff} = 7$ the AIC^*_{diff} and X^2_{diff} test lead to similar conclusions; (c) for $df_{diff} > 7$, the AIC^*_{diff} rule becomes increasingly more conservative than the X^2_{diff} test (i.e., more likely to accept the more restrictive model). Demonstrating this point from a slightly different perspective, Bozdogan (1987) noted that the probability of a type I error, accepting the more complex model when the more parsimonious model was true, varied from .15 for $df_{diff} = 1$, to .05 for $df_{diff} = 7$, to .001 for $df_{diff} = 28$. Hence, it is clear that the AIC^*_{diff} and X^2_{diff} tests will only lead to substantially different results when the df_{diff} is moderately large, thus explaining why the two decision rules lead to similar decisions in some applications. If one questions the practical usefulness of the X^2 for evaluating a priori models, then one must also question the practical usefulness of the AIC* when df_{diff} is not large. A similar logic applies to the Ck and SC*, except that these tests tend to be more conservative than the AIC*.

To further illustrate the behavior of these indices, consider Models M3, M7, and M8 for the CSIM data (Appendix 2B): M7 is the most parsimonious true model, M3 is a less parsimonious true model, and M8 is false but more parsimonious than M3 or M7. The three indices lead to the selection of M7 over M3, but differ in comparisons based on M3 vs. M8 and M7 vs. M8. For the comparison of M3 vs. M8, AIC*, Ck and SC* all lead to the selection of the more parsimonious false model for small N and the less parsimonious true model for the large N. For intermediate N, AIC* and Ck lead to the selection of the less parsimonious true model, whereas SC* leads to the selection of the false model. For the comparison of M7 vs. M8, AIC* leads to the selection of the correct model at all Ns, Ck leads to the selection of the correct model for all but the smallest N, and SC* leads the selection of the correct model for all but the two smallest Ns. This behavior of the indices is consistent with the relative size of the penalty function associated with each index and the influence of N noted earlier (also see Browne & Cudeck, 1989; Cudeck & Browne, 1983). Whereas there are systematic differences between these indices, the results indicate that all three indices may over-penalize model complexity in relation to criteria established here, particularly when N is small.

Summary of penalty functions It is useful to distinguish between some characteristics of penalties for model complexity in indices considered here. The penalty functions for RMS, AGFI*, and TLI do not vary with N but do vary with goodness of fit and they provide no penalty for model complexity when Dk = 0. The penalty for model complexity in the parsimony indices does not vary with N so that parsimony indices based on S

provide unbiased estimates of values based on Σ so long as the parsimony index is based on an unbiased index (e.g., RNI instead of BBI). The penalty functions in the AIC*, SC*, and Ck reflect estimation discrepancy that is a systematically descreasing function of N and is independent of approximation discrepancy. For a sufficiently large N, the penalty functions in AIC*, SC*, and Ck approach zero so that there is no penalty for model complexity and no compromise between goodness of fit and model complexity. Because the penalties do vary as a function of N, AIC*, SC*, and Ck based on S provide biased estimates of approximation discrepancy based on Σ. The indices have been recommended as an alternative to the traditional X^2 test of statistical significance, but the indices suffer many of the same limitations as the X^2 in some applications.

Recomendations For the Use of Goodness of Fit Indices

Based on results of this study, I offer the following recommendations for evaluating goodness of fit. Communication in this area can be improved by making clear distinctions between approximation discrepancy, estimation discrepancy, parsimony penalties for model complexity applied to indices of approximation discrepancy that are independent of N, and estimation penalities that reflect estimation discrepancy and vary inversely with N. In relation to criteria specified earlier, I found support for 6 indices: Dk, Mc, GFI*, AGFI*, TLI, and RNI. The results provided no clear basis for choosing among these indices, but apparently important distinctions warrant consideration.

Dk, Mc, and GFI* are stand-alone indices that are monotonically related to model complexity. They are particularly uUseful when the intent is to compare the fits of a relatively small number of a priori models that have a well-defined theoretical rationale. Because these indices are monotonically related to model complexity, caution is needed in evaluating and selecting alternative models. For example, if the a priori models are nested, then each of these indices will automatically select the most complex model. Hence evaluation of the alternative models will require the subjective combination of theory, substantive issues, interpretability, and parsimony as well as the application of these indices. Also, because none of these indices has been widely applied, there are no well established guidelines for evaluating what values constitute a "good" fit.

RNI is an incremental index that is also monotonically related to model complexity. When used in conjunction with a suitably defined null model, RNI scales goodness of fit along a 0-to-1 continuum. Sampling fluctuations will ocassionaly result in RNI values outside the 0-to-1 continuum, but this apparently presents no problems in the interpretation; RNIs > 1.0 clearly reflect a good fit whereas RNIs < 0 refect a bad fit. An apparent advantage of the RNI is that experience with incremental fit indices suggests approximate guidelines for what values (e.g., RNIs > .90) constitute an "acceptable fit," although there is apparently no rigorous defense of such guidelines. It is also possible to define RNI in terms of models of intermediate complexity instead of the null and the saturated models. This is not particularly useful, however, since the RNIs are strictly additive. That is, the RNI based on two models of intermediate complexity is merely the difference in RNIs of the same

two models that are defined in relation to the saturated and null models. Also, guidelines for interpreting acceptable fits are no longer appropriate if the RNI is not defined in relation to the null and saturated model.

Dk for the null model is a constant for any given sample so that RNI is a linear transformation of Dk in any particular sample and the difference in RNIs for any two models is a linear transformation of the difference in Dks. Hence the main advantage for the use of RNI apparently rests on the suggestion that the absolute values of RNIs are easier to interpret than those of Dk. Given the wide use of implicit guidelines such as the "greater than .90 rule" prompted by Bentler and Bonett's (1980)[6] seminal article , it is disappointing that apparently so little research has evaluated this guideline. Thus, whereas I reccommend the use of incremental fit indices, this is based in large part on the unsubstantiated belief that guidelines used to evaluate their absolute size are meaningful. Otherwise, there may be no advantage of the incremental indices over Dk or suitable transformations of it such as Mc.

TLI and AGFI* differ from the other recommended indices in that they contain a parsimony penalty for model complexity. The size of the penalty varies inversely with approximation discrepancy so that it typically has limited effect on models with small Dks or "acceptable" RNIs greater than .9. These indices have the disadvantage of being undefined for the saturated model and may be unstable when the df is close to zero or when the sample size is very small. The TLI is an incremental index in that it varies along a 0-to-1 scale (except for sampling fluctuations) in which the endpoints are defined by the fit of a null model and a "perfect" fit. The parsimony penalty may, however, complicate the typical interpretation of incremental fit indices like RNI. Thus, for example, the TLI may be maximized for a model of intermediate complexity so that TLIs are not strictly additive. The TLI and AGFI* indices are apparently most useful in comparing the fit of a priori models that differ substantially in the number of estimated parameters. Particularly when df or N are small, the indices should be used in combination with other indices such as Dk or RNI.

In addition to the six recommended indices, it is also useful to comment on the parsimony indices and the information criterion indices. Whereas both are designed to penalize for model complexity, their rationales are very different. The parsimony indicies incorporate a parsimony penalty for model complexity that does not vary with N, whereas the information criterion indices incorporate an estimation penalty that is a monotonically decreasing function of N to control for sampling fluctuations. Throughout my discussion I have emphasized that model parsimony is a very important criterion that should be considered in relation to model evaluation and selection. Whereas this is the explicit intent of the parsimony indices, the rigid form of combining information on goodness of fit and parsimony embodied in these indices seems to lack a clear rationale and in some instances may be inappropriate. Hence, I recommend that goodness of fit and parsimony should be considered separately and that information from these two criteria should be combined subjectively within the context of a particular study. It would be useful, however, for researchers to report parsimony ratios along with other fit indices to facilitate integration of parsimony into the evaluation and selection of models.

Cudeck and Browne (1983) rejected the usefulness of the X^2 test for the evaluation of a priori covariance structure models and proposed the use of information criterion indices as an alternative indicator of fit. A potential limitation of these indices is that they employ a decision rule that has a simlar logic to the X^2 test and frequently result in the selection of similar models when comparing nested models that differ only moderately in terms of complexity. Nevertheless, when N is small and the number of estimated parameters is large, it is important that sampling fluctuation is considered in the evaluation of models. This could be accomplished by presenting separate estimates of approximation discrepancy and estimation discrepancy implicit in the information criterion indices. Alternatively, it may be more useful to present confidence intervals about unbiased estimates of approximation discrepancy such as Dk as recommended by Steiger and Lind (1980; Steiger, 1990). In either case, I do not recommend that the estimates of approximation discrepancy and estimation discrepancy are combined in a single index that confounds the two sources of information.

Model evaluation inevitably requires a subjective integration of statistical and nonstatistical criteria, but an unbiased estimate of how well a model would fit population data should be one of these considerations. Other information such as parsimony, sampling fluctuations, comparative fit, interpretability, substantive issues and theory should all be considered. Nevertheless, I remain skeptical about indices that seek to combine (and confound) different sources of information into a single decision rule that is used mechanisitcally in all situations. I prefer to retain a degree of flexibility and subjectivity. Instead of seeking the "one" index to evaluate all models in all situations, I recommend that researchers focus on what sources of information are most relevant in their particular situation.

Footnotes

1 -- Bollen (1990, p. 256) noted the distinction between the questions of (1) "whether the means of a sampling distribution for a fit index are associated with sample size" and (2) "whether N directly enters into the calculation of a fit measure," and criticized Marsh, Balla and McDonald (1988) for not recognizing this distinction. Here I focus on the first question, noting that the second question is not very useful except as it relates to the first because fit indices can typically be expressed in mathematically equivalent ways that do or do not contain N (e.g., any fit index that can be expressed as a function of the maximum likelihood fitting function can also be expressed as a function of X^2/N).

2 -- Parsimony and complexity as inferred from the parsimony ratio differs from the traditional factor analysis concept of "simple structure." For independent clusters models in which each indicator loads on a single factor, the parsimony ratio is a monotonically increasing function of the number of factors and the average number of indicators per factor. For example, in a "simple" independent clusters model in which each factor has four indicators, the parsimony ratios are .33, .68, .77, .82, .84, and .89 for models with 1, 2, 3, 4, 5, and 10 factors respectively (see Table 7 and subdequent discussion of parsimony indices).

3 -- It is interesting to note that if RMS were defined as Dk/df instead of its square root, then the TLI is the difference between RMS for a target model and an null model.

4 -- The parsimony ratios for the models selected by the Ck in the Cudeck and Henly (1991) study were .36, .14, .14, and .05 for models based on Ns of 100, 300, 500, and 1200 respectively. Whereas standards for evaluating parsimony are not well defined, these appear to be much less parsimonious than the a priori models typically encountered in practice.

5 -- It is interesting to note that for the comparison of nested models, the AIC* decision rule leads to the same recommendation as guideline to accept models that have a X^2/df ratio less than 2.

6 -- Although I attribute the "greater than .9 rule" to Bentler and Bonett (1980), they actually stated that "since the scale of the fit indices is not necessarily easy to interpret ... experience will be required to establish values of the indices associated with various degrees of meaningfulness" and "in our experience, models with overall fit indices of less than .9 can usually be improved substantially" (p. 600). Bentler (1990) says even less about what values constitute an acceptable fit except to note that different rules of thumb may have to be applied to incremental indices like the TLI that incorporate an esimtation penalty.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716-723.

Akaike, H. (1987). Factor analysis and AIC. Psychometrika, 52, 317-332..

Bentler, P. M. (1990). Comparative fit indices in structural models. Psychological Bulletin, .

Bentler, P. M. & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Bollen, K. A. (1986). Sample size and Bentler and Bonett's nonnormed fit index. Psychometrika, 51, 375-377.

Bollen, K. A. (1986). Overall fit in covariance structure models: Two types of sample size effects. Psychological Bulletin, 107, 256-259.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 345-370.

Browne, M. W. & Cudeck, R. (1983). Single-sample cross-validation indices for covariance structures. Multivariate Behavioral Research, 24, 445-455.

Cox, D. R., & Hinkley, D. V. (1974). Theoretical statistics, London: Chapman and Hall.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. Multivariate Behavioral Research, 18, 147-167.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification.. Psychological Bulletin, 109, 512-519.

Hoelter, J. W. (1983) The analysis of covariance structures: Goodness- of-fit indices. Sociological Methods & Research, 11, 325-344.

Hull, C. H., & Nie, N. H. (1981). SPSS update 7-9, New York: McGraw-Hill.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). Causal analysis. Assumptions, models, and data. Beverly Hills, CA: Sage.

Joreskog, K. G. & Sorbom, D. (1981). LISREL V: Analysis of Linear Structural Relations By the Method of Maximum Likelihood, Chicago: International Educational Services.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. Multivariate Behavioral Research, 25, 137-155.

Lawley, D. N., & Maxwell, A. E. (1971). Factor analysis as a statistical method, New York: American Elsevier.

MacCallum. (1990). The need for alternative measures of fit in covariance structure modeling. Multivariate Behavioral Research, 25, 157-162.

Marsh, H. W. (1987). The factorial invariance of responses by males and females to a multidimensional self-concept instrument: Substantive and methodological issues. Multivariate Behavioral Research, 22, 457-480.

Marsh, H. W., Balla, J. R. & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 102, 391-410..

McDonald, R. P. (1985). Factor analysis and related methods, Hillsdale, NJ: Erlbaum.

McDonald, R. P, & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness-of-fit. Psychological Bulletin, 107, 247-255.

McDonald, R. P. (1989). An Index of Goodness-of-fit based on noncentrality. Journal of Classification, 6, 97-103.

McArdle, J. J. (1986). Latent variable growth within behavior genetic models. Behavior Genetics, 16, 163-200.

Mulaik, S. A. (October, 1991). Clarifying Misconceptions About Parsimony Adjustments of Goodness of Fit Indices. Paper presented at the 1991 Annual Meeting of the Society of Multivariate Experimental Psychology.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105, 430-445.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. Psychometrika, 54, 131-151.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis: A unified approach. Psychometrika, 50, 83-90.

Schwartz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

Steiger, J. H. (1989). EzPATH: A supplementary module for SYSTAT and SYSGRAPH. Evanston, IL: SYSTAT, Inc.

Steiger, J. (1990). Structure model evaluation and modification: An interval estimation approach. Multivariate Behavioral Research, 25, 173-180.

Steiger, J. H., & Lind, J. M. (May, 1980). Statistically-based tests for the number of common factors. Paper presented at the Psychometrika Society Meeting, Iowa City.

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. Child Development, 58, 134-146.

Tanaka, J. S., & Huba, G. J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. British Journal of Mathematical and Statistical Psychology, 46, 621-635.

Tucker, L. R. & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society, 54, 426-482.

Table 1

Population Parameters Used to Generate the SSIM and CSIM Population Data Structures

SSIM Population Data Structure

	Factor Loadings			
	F1	F2	F3	Uniquenesses
x1	.6	0	0	1.0
x2	.7	0	0	1.0
x3	.8	0	0	1.0
x4	0	.6	0	1.0
x5	0	.7	0	1.0
x6	0	.8	0	1.0
x7	0	0	.6	1.0
x8	0	0	.7	1.0
x9	0	0	.9	1.0

Factor Variance/Covariances

F1	1.04		
F2	.08	1.16	
F3	.12	.24	1.36

CSIM Population Data Structure

	Factor Loadings			
	F1	F2	F3	Uniquenesses
x1	.6	.2	0	1.0
x2	.7	0	.2	1.0
x3	.8	0	0	1.0
x4	.2	.6	0	1.0
x5	0	.7	.2	1.0
x6	0	.8	0	1.0
x7	.2	0	.6	1.0
x8	0	.2	.7	1.0
x9	0	0	.9	1.0

Factor Variance/Covariances

F1	1.04		
F2	.08	1.16	
F3	.12	.24	1.36

Table 2

Pattern of Factor Loadings For Eight Models (M1 - M8) Used to Fit Covariance Matrices Based on 9 Measured Variables (x1 - x9)

	M1	M2		M3			M4			M5			M6			M7[a]			M8[a]		
	F1	F1	F2	F1	F2	F3	F1	F2	F2	F1	F2	F2	F1	F2	F2	F1	F2	F2	F1	F2	F2
	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
x1	1	1	0	1	0	0	1	1	0	1	0	0	1	0	0	1	1	0	1	0	0
x2	1	1	1	1	1	0	1	0	0	1	0	1	1	0	0	1	0	1	1	0	0
x3	1	1	1	1	1	1	1	0	0	1	0	0	1	1	0	1	0	0	1	0	0
x4	1	1	1	1	1	1	1	1	0	0	1	0	0	1	0	1	1	0	0	1	0
x5	1	1	1	1	1	1	0	1	0	0	1	1	0	1	0	0	1	1	0	1	0
x6	1	1	1	1	1	1	0	1	0	0	1	0	0	1	1	0	1	0	0	1	0
x7	1	1	1	1	1	1	1	0	1	0	0	1	0	0	1	1	0	1	0	0	1
x8	1	1	1	1	1	1	0	0	1	0	1	1	0	0	1	0	1	1	0	0	1
x9	1	1	1	1	1	1	0	0	1	0	0	1	1	0	1	0	0	1	0	0	1

Note. 1s represent factor loadings that were freely estimated whereas 0s represent factor loadings that were fixed to be 0. In addition, each model had 9 estimated (uncorrelated) uniquenesses corresponding to the 9 measured variables. All models also had factor variances fixed to 1, and for the last 5 models factor covariances were freely estimated. M1 = 1 Unrestricted Factor model (18 estimated parameters); M2 = 2 Unrestricted Factors model (26 estimated parameters); M3 = 3 Unrestricted Factors model (33 estimated parameters); M4 = 3 Factors model (24 estimated parameters); M5 = second 3-Factor model (24 estimated parameters); M6 = third 3 Factor model (24 estimated parameters); M7 = 3 Complex Factors model (27 estimated parameters); M8 = 3 Simple Factors model (21 estimated parameters).

[a] The M7 model is the "correct" model for the complex simulated (CSIM) population data structure in that all non-zero parameters in the known population are freely estimated and all zero parameters in the population are fixed to be zero (see Appendix 2). Similarly, the M8 model is the "correct" model for the simple simulated (SSIM) population data structure.

BEST COPY AVAILABLE

Table 3

Description of A Priori Contrasts To Be Tested.

	Models[a]								Predictions For[b]:	
	M1	M2	M3	M4	M5	M6	M7	M8	CSIM Data	SSIM Data
Contrast 1	-1	+1	0	0	0	0	0	0	M2 > M1	M2 > M1
Contrast 2	0	-1	+1	0	0	0	0	0	M3 > M2	M3 > M2
Contrast 3	0	0	+1	0	0	0	-1	0	M3 = M7	M3 = M7
Contrast 4	0	0	+1	0	0	0	0	-1	M3 > M8	M3 = M8
Contrast 5	0	0	0	+1	+1	-2	0	0	M4 ,M5 > M6	M4 ,M5 = M6
Contrast 6	0	0	0	-1	-1	0	+2	0	M7 > M4 ,M5	M7 = M4 ,M5
Contrast 7	0	0	0	+1	+1	0	0	-2	M4 ,M5 > M8	M4 ,M5 = M8
Contrast 8	0	0	0	0	0	+1	0	-1	M6 = M8	M6 = M8
Contrast 9	0	0	0	0	0	0	+1	-1	M7 > M8	M7 = M8

Note. The simple simulated (SSIM) population data structure has the same pattern of zero and nonzero (known) parameters as M8, whereas the complex simulated (CSIM) population data structure has the same pattern of zero and nonzero (known) parameters as M7 (see Tables 1 and 2).

[a] See Table 2 for a description of the models. [b] For predictions represented by > signs, models on the left side of the > signs should be better able to fit the data. For all but one prediction represented by = signs, models on the left side of the = signs have more freely estimated parameters known to be zero in the population and thus provide a test of penalty functions imposed by some indices. For just contrast 5 for the SSIM data, models on both sides of the = sign have the same number of parameters.

Table 4

The Effect of Model and Sample Size on Fit Indices for SSIM Data: Effect Sizes (Etas and rs) and A Priori Contrasts

a		b				Inter-	c							
Index		Model	Size			action	A Priori Contrasts							
									d	d	d	d	d	d	d
		Eta	Eta	r1	r2	Eta	1	2	3	4	5	6	7	8	9
Stand-alone indices															
1	FF	.82**	.47**	-.43	-.31	.14**	+**	+**	+**	+**	+	+**	+**	+**	+**
2	LHR	.81**	.50**	.45	.33	.14**	+**	+**	+**	+**	+	+**	+*	+**	+**
3	X2	.65**	.36**	.36	.33	.33**	+**	+**	+**	+**	+	+**	+*	+**	+**
4	X2/df	.64**	.38**	.38	.34	.66**	+**	+**	+	+	+	+	+	-	+
5	RMR	.84**	.41**	-.31	-.39	.21**	+**	+**	+**	+**	-	+**	+**	+**	+**
6	RMS	.90**	.13*	-.12	-.09	.16**	+**	+**	-	+	+	+	+	-	+
7	GFI	.81**	.50**	.45	.33	.16**	+**	+**	+**	+**	+	+**	+**	+**	+**
8	AGFI	.72**	.59**	.54	.40	.15**	+**	+**	+	+*	+	+*	+	+	+*
9	GFI*	.93**	.02	-.01	-.01	.10	+**	+**	-	+	+	+	+*	-	+*
10	AGFI*	.90**	.03	-.02	-.02	.12*	+**	+**	-	+	+	+	+	-	+
11	AIC*	.40**	.90**	-.81	-.60	.06*	+**	+**	-**	-**	+	-**	-**	-**	-**
12	SC*	.23**	.96**	-.90	-.68	.09**	+**	+**	-**	-**	+	-**	-**	-**	-**
13	Ck	.33**	.93**	-.82	-.59	.07*	+**	+**	-**	-**	+	-**	-**	-**	-**
14	CN	.35**	.67**	.67	.60	.39**	+**	+**	+**	+**	+	+*	+	+	+
15	Dk	.93**	.02	.01	.01	.10	+**	+**	-	+	+	+	+*	-	+*
16	MC	.93**	.02	.01	-.01	.10	+**	+**	-	+	+	+	+*	-	+*
Type-1 incremental indices															
17	FFI1	.91**	.26**	.25	.20	.18**	+**	+**	+**	+**	+	+**	+**	+**	+**
18	BBI	.91**	.26**	.25	.19	.18**	+**	+**	+**	+**	+	+**	+**	+**	+**
Type-2 incremental indices															
19	TLI	.92**	.07	-.05	-.04	.14**	+**	+**	-	+	+	+	+	-	+
20	RNI	.95**	.05	-.03	-.02	.12**	+**	+**	-	+	+	+	+	-	+
Parsimony Indices															
21	PBBI	.91**	.25**	.25	.19	.19**	+**	-**	-**	-**	+	-**	-**	-**	-**
22	PRNI	.95**	.05*	-.04	-.03	.12**	+**	-**	-**	-**	+	-**	-**	-**	-**

Note.. Results are based on a series of 8 (Models) by 6 (Sample Sizes) ANOVAs conducted on the simple simulated population data structures.

* $p < .05$; ** $p < .01$.

[a] All 22 indices are defined in Appendix 1. [b] Eta is the linear and nonlinear effects of sample size (N), r1 is the linear effect of the log sample size (sample sizes are log spaced in this study), and r2 is the linear effect of sample size. [c] For each of the a priori contrasts a + sign indicates that the best fit was obtained for the model posited to fit the best, or for the model with the greatest number of parameters when the contrasted models were posited to fit equally well (see Table 3 for a description of the a priori contrasts). [d] These contrasts are between models that are equally able to fit the data.

Table 5
Effect of Sample Size and Model on Fit Indices for SSIM Data For Models 3-8

Index[a]	Model[b]		Size[c]			Interaction
	Eta	Parsimony	Eta	r1	r2	Eta
Stand-alone indices						
1 FF	.18**	+	.89**	-.79	-.53	.20**
2 LHR	.18**	+	.90**	.81	.60	.19**
3 X2	.52**	+	.07	-.02	-.01	.07
4 X2/df	.05	+	.08	-.02	-.01	.10
5 RMR	.29**	+	.89**	-.86	-.70	.16**
6 RMS	.05	-	.46**	-.43	-.34	.09
7 GFI	.21**	+	.89**	.81	.60	.21**
8 AGFI	.05	+	.93**	.84	.63	.06
9 GFI*	.07	+	.11	.06	.03	.05
10 AGFI*	.06	+	.06	.04	.02	.13
11 AIC*	.05**	-	.99**	-.66	-.89	.04*
12 SC*	.11**	-	.99**	-.92	-.70	.08**
13 Ck	.05**	-	.99**	-.88	-.63	.06**
14 CN	.06**	+	.82**	.74	.82	.07*
15 Dk	.07	+	.12	-.07	-.04	.14*
16 Mc	.07	+	.10	.05	.03	.14
Type-1 incremental indices						
17 FFI1	.21**	+	.84**	.80	.63	.19**
18 BBI	.21**	+	.84**	.80	.63	.19**
Type-2 incremental indices						
19 TLI	.06	+	.06	.03	.01	.03
20 RNI	.07	+	.11	.07	.04	.13
Parsimony Indices						
21 PBBI	.89**	-	.38**	.37	.29	.14**
22 PRNI	.97**	-	.03	.02	.01	.03

Note.. Results are based on a series of 6 (Models) by 6 (Sample Sizes) ANOVAs conducted on the simple simulated population data structure. For purposes of these analyses only the three-factor models, all of which are true models that are able to fit the data, were included.

* $p < .05$; ** $p < .01$.

[a] All 22 indices are defined in Appendix 1. [b] Because all models are equally able to fit the data, differences between models are a test of the relations between each model and model complexity. Under the Parsimony column a + indicates that fit improved with the addition freely estimated parameters known to be zero in the population and a - indicates that fit was poorer with the addition these freely estimated parameters. [c] Eta is the linear and nonlinear effects of sample size (N) , r1 is the linear effect of the log sample size (sample sizes are log spaced in this study), and r2 is the linear effect of sample size.

Table 6

The Effect of Model and Sample Size on Fit Indices for CSIM Data: Effect Sizes (Etas and rs) and A Priori Contrasts

Index[a]	Model	Size[b]			Interaction	A Priori Contrasts[c]								
								[d]					[d]	
	Eta	Eta	r1	r2	Eta	1	2	3	4	5	6	7	8	9
Stand-alone indices														
1 FF	.73**	.57**	-.51	-.36	.16**	+**	+**	+**	+**	+*	+**	+**	+**	+**
2 LHR	.74**	.58**	.52	.38	.14**	+**	+**	+**	+**	+**	+**	+**	+**	+**
3 X2	.60**	.49**	.44	.49	.62**	+**	+**	+**	+**	+**	+**	+**	+**	+**
4 X2/df	.57**	.53**	.47	.53	.61**	+**	+**	+	+**	+**	+**	+**	+	+**
5 RMR	.80**	.46**	-.43	-.32	.19**	+**	+**	+**	+**	+*	+**	+**	+**	+**
6 RMS	.85**	.07	-.06	-.04	.14**	+**	+**	+	+**	+*	+**	+**	+	+**
7 GFI	.74**	.57**	.50	.36	.16**	+**	+**	+**	+**	+	+**	+**	+**	+**
8 AGFI	.61**	.68**	.61	.44	.15**	+**	+**	+	+**	+	+**	+**	+**	+**
9 GFI*	.88**	.05	.02	.00	.11**	+**	+**	+	+**	+**	+**	+**	+	+**
10 AGFI*	.84**	.04	.02	.00	.13**	+**	+**	+	+**	+**	+**	+**	+	+**
11 AIC*	.30**	.94**	-.84	-.62	.05**	+**	+**	-**	+**	+**	+**	+*	+	+**
12 SC*	.16**	.98**	-.91	-.69	.09**	+**	+**	-**	-**	+**	-	-*	-**	-**
13 Ck	.24**	.96**	-.84	-.61	.07**	+**	+**	-**	+	+**	+**	+**	-	+**
14 CN	.53**	.48**	.46	.48	.61**	+**	+**	+*	+**	+**	+**	+**	+*	+**
15 Dk	.88**	.06	-.02	-.01	.11*	+**	+**	+	+**	+**	+**	+**	+	+**
16 MC	.88**	.05	.02	.00	.12**	+**	+**	+	+**	+**	+**	+**	+	+**
Type-1 incremental indices														
17 FFI1	.87**	.37**	.35	.27	.15**	+**	+**	+**	+**	+**	+**	+**	+**	+**
18 BBI	.87**	.37**	.35	.27	.15**	+**	+**	+**	+**	+**	+**	+**	+**	+**
Type-2 incremental indices														
19 TLI	.85**	.07	-.04	-.03	.12**	+**	+**	+	+**	+*	+**	+*	+	+**
20 RNI	.90**	.04	-.03	-.02	.10*	+**	+**	+	+**	+*	+**	+**	+	+**
Parsimony indices														
21 PBBI	.87**	.34**	.33	.25	.16**	-	-**	-**	-**	+**	-**	-**	-**	-**
22 PRNI	.91**	.03	-.03	-.02	.09	-	-**	-**	-**	+*	-**	-**	-**	-**

Note. Results are based on a series of 8 (Models) by 6 (Sample Sizes) ANOVAs conducted on the complex simulated (CSIM) population data structure.

* $p < .05$; ** $p < .01$.

[a] All 22 indices are defined in Appendix 1. [b] Eta is the linear and nonlinear effects of sample size (N) , r1 is the linear effect of the log sample size (sample sizes are log spaced in this study), and r2 is the linear effect of sample size. [c] For each of the a priori contrasts a + sign indicates that the best fit was obtained for the model posited to fit the best, or for the model with the greatest number of freely estimated parameters known to be zero in the population when the contrasted models were posited to fit equally well (see Table 3 for a description of the a priori contrasts). [d] These contrasts are between models that are equally able to fit the data.

BEST COPY AVAILABLE

Table 7

Parsimony Ratios (PR) and Minimum Value of the Unbiased Relative Fit Index (RNI) required to Achieve the Minimally Acceptable Parsimony Index of .80 as a Function of the Number of Factors and Number of Indicators Per Factor

Number of Indicators per factor	Number of Factors											
	1		2		3		4		5		10	
	PR	Min RNI	PR	Min RNI	PR	Min RNI	PR	Min RNI	PR	Min RNI	PR	Min RNI
3	0	---	.53	---	.67	---	.73	---	.76	---	.83	.97
4	.33	---	.68	---	.77	---	.82	.98	.84	.95	.89	.90
5	.50	---	.76	---	.83	.96	.86	.93	.88	.91	.92	.87
6	.60	---	.80	1.00	.86	.93	.89	.90	.91	.88	.94	.85
7	.67	---	.84	.95	.89	.90	.91	.88	.92	.87	.95	.84
8	.71	---	.86	.93	.90	.89	.92	.87	.94	.85	.96	.83

APPENDIX I

Descriptions of the 22 Goodness of Fit Indices Used in This Study

---

For present purposes I define: S = sample covariance matrix; Σ = true population covariance matrix (estimated by S); E = fitted covariance matrix which will not Σ when the model being fit is misspecified; and Σ_E = the population covariance matrix that is estimated by E. Four types of fit indices are considered here. Stand alone indices are based on results of just the hypothesized model. Two forms of incremental indices, called type-1 and type-2 for present purposes, are based on differences in fit between a hypothesized model and a null model. Parsimony indices are constructed by multiplying any goodness of fit index by the parsimony ratio as a penalty for model complexity. All fit indices described here a sampling quantities.

I. Absolute, Stand-alone Indices.

The maximum likelihood fitting function (FF) and the scaled likelihood ratio (LHR). Although not typically presented as fit indices (but see Cudeck & Browne, 1983), the FF and LHR are the basis for the X^2 test statistic and most other fit indices. The FF has a minimum value of 0 when E = S, but does not have an upper bound. The scaled LHR has a maximum value of 1.0 when E = S and a minimum value of zero. The FF and LHR are defined as:

(1) $FF = X^2/(N)$,

(2) $LHR = e^{-1/2\ FF}$.

X^2 and X^2/df Ratio. These two indices continue to be the most frequently used indices. The X^2 for a misspecified model varies directly with sample size, but the X^2 for a true model does not. In CFA the df does not vary with the sample size, so that the effect of sample size on the X^2/df must necessarily be the same as for the X^2. For alternative models of the same data, increasing the number of parameters necessarily results in a better (i.e., lower) X^2. Because the X^2/df ratio incorporates a penalty function for using more parameters, it may be poorer if additional parameters result in little improvement in X^2. They are defined as:

(3) $X^2 = tr\ (E^{-1}\ S - I) - \log |\ E^{-1}\ S\ | = (N)\ FF$,

(4) $X^2/df = ((N)/df)\ FF$.

LISREL's root mean square residual (RMR). Joreskog and Sorbom (1981, p. I.41) define the RMR as the square root of the mean of squared residuals in S and E. When S and E are based on correlation matrices RMR is strictly bounded by 0 and 1. For covariance matrices RMR still has a lower-bound of zero but does not have an upper bound. Thus RMR must be interpreted in relation to the size of the variances and covariances of the measured variables, and cannot be compared across applications based on different variables. RMR is defined as:

(5) $RMR = [\ 2\ \Sigma_{ij}\ (s_{ij} - e_{ij})^2\ /(p \times (p+1))]^{1/2}$.

where s_{ij} and e_{ij} are elements in S and E

Steiger-Lind Adjusted root mean square index (RMS). Steiger and Lind (1980; Steiger, 1989) noted Dk (see below) was not in the metric of the original parameters and may not compensate for model complexity. Steiger and Lind (1980) proposed the RMS index to address these problems that is defined as:

(6)[a] $RMS = (Dk/df_t)^{1/2}$ if Dk is non-negative,

RMS = 0 if Dk is negative.

where Dk is defined as below. Steiger (1990, p. 81) suggested that: "Values below .10 indicate a good fit, values below .05 a very good fit. Point estimates below .01 indicate an outstanding fit, and are seldom obtained."

LISREL's goodness-of-fit (GFI) and adjusted GFI (AGFI). Joreskog and Sorbom (1981; also see Tanaka & Huba, 1986) describe the GFI and AGFI as computed by LISREL. They state that GFI is "a measure of the relative amount of variances and covariances jointly accounted for by the model" and assert that "unlike X^2, GFI is independent of the sample size" while AGFI "corresponds to using mean squares instead of total sums of squares" (Joreskog & Sorbom, 1981, p. I. 40-41). Thus AGFI incorporates a penalty function for additional parameters. Joreskog and Sorbom suggest that GFI and AGFI will generally fall between 0 and 1, but that it is possible for them to be negative. They are defined as:

(7) $GFI = 1 - [(tr (E^{-1} x S - I))^2/(tr E^{-1} S)^2]$,

(8) $AGFI = 1 - [p x (p+1)/2df] x (1 - GFI)$.

Steiger's Estimate of Population Quantities of GFI and AGFI (GFI* and AGFI*)

Steiger (1989; also see McDonald & Marsh, 1990) conjectured that GFI is a biased estimate of the corresponding population quantity. Based on earlier work by Browne (1974) and by Tanaka and Huba (1988), Steiger (1989) proved that the population equivalent of the GFI can be expresses as a simple function of the populaton noncentrality index. Using McDonald's Dk (see below) as one consistent estimator of the population noncentrality index, a consistent estimate GFI* can be written as:

(9)[a] $GFI^* = p/(2Dk + p)$.

where Dk is defined as below. Steiger (1989) noted that GFI* continues to improve as more parameters are added and thus does not take into account model complexity. In order to address this issue he defined an adjusted GFI* as:

(10)[a] $AGFI^* = 1 - [p x (p+1)/2df] x (1 - GFI^*)$.

Information Criterion. Akaike (1974, 1987) and Schwartz (1978) each proposed fit indices that incorporate penalty functions based on the number of parameters for comparing alternative models. Cudeck and Browne (1983, p. 154) and Steiger (1989) considered rescaled versions of these models. Steiger (1989, p.87) noted that "we rescale the criterion (without affecting the decisions it indicates) so that it remained more stable across differing sample sizes." proposed rescaled versions of these indices expressed in terms of FF. Because this is considered a desirable feature in the context of the present investigation, I consider the rescaled versions of these indices defined as:

(11) $AIC^* = FF + 2K / N$,

(12) $SC^* = FF + (K x ln(N)) / N$

where K = the number of parameters to be estimated (see Bozdogan, 1987, for more complete derivations).Browne-Cudeck (1989) Single Sample Cross-Validation Index (Ck). Cudeck and Browne (1983) described a cross-validation procedure for testing alternative models. They also considered the correspondence between their approach to the information criteria AIC* and SC*. Browne and Cudeck (1989) extended their earlier research, developing a better approximation to the cross-validation coefficient based on a single sample that is defined as:

(13) $Ck = FF + 2K/(N - p - 2)$

Critical N (CN). Hoelter (1983, p. 528) argued that "rather than ignoring or completely neutralizing sample size I can estimate the size that a sample must reach in order to accept the fit of a given model on a statistical basis. This estimate, referred to here as 'critical N' (CN), allows one to assess the fit of a model relative to identical hypothetical models estimated with different sample sizes." Hoelter cautioned that no firm basis could be offered as to what constituted an adequate fit, but he suggested that a value of 200 was a reasonable starting point for suggesting that differences between the model and data may be unimportant. In practice the usefulness of CN would rest on the assumption that its value is independent of sample size. It is defines as:

(14) $CN = [[z_{crit} + (2 \times df - 1)^{1/2}]^2/[2 \times X^2/(N)]] + 1.$

where z_{crit} = the critical value from a normal curve table for a given probability level -- 1.96 in the present investigation.

Noncentrality Parameter Indices. McDonald (1990) notes that a problem with the AIC, as with many other fit indices, is that the value of the index and model selection based on it are dependent on sample size. He considered a noncentrality index that is based on similar formulations as the AIC but with a slightly different derivation. McDonald suggested Wald's (1943) noncentrality parameter (also see very similar suggestions by Steiger and Lind, 1980, and subsequent discussion by Steiger, 1989), rescaled to be independent of sample size, as an index of fit, estimated by:

(15)[b] $Dk = FF - df/N = AIC - (2K/N) - df/N.$

A nontechnical interpretation of the noncentrality parameter is as a suitably weighted sum of squares of discrepancies between the parameters of the unrestricted model yielding Σ and the parameters of the approximating model yielding Σ_E. McDonald further noted that DK could be transformed to yield Mc, a measure of centrality that is a consistent estimator of the population discrepancy function, scaled to be independent of sample size. It is in the metric of LHR (2) and bears the same relation to Dk (15) as the LHR (2) does to FF (1) as noted above. Mc is scaled to lie on the interval zero to unity with unity representing a perfect fit, though sampling error may produce values outside of this range. It is defined as:

(16)[b] $Mc = e^{DK/2}$

II. Relative, Type-1 Incremental Fit Indices.

Bentler and Bonett (1980) proposed that valuable information could be obtained by comparing the ability of nested models to fit the same data. In the case of CFA it may be useful to compare the fit of the proposed target model with the fit of a null model in which all the p variables are assumed to be uncorrelated. (It should be noted that in general models for the analysis of covariance structures the null model is not the only more restrictive model that could be considered as a baseline model.) If the fit of a null model is reasonable, because the sample size is small or because the measured variables are relatively uncorrelated, then the difference in fit between the null and target models will be small. However, if the fit of the null model is reasonable then there is little covariance to explain and no basis of support for the target model even if it also fits the data. Bentler and Bonett specifically stated that these indices are useful for comparing the fit of a particular model across samples that have unequal sizes. They cautioned that the absolute value of these indices may be difficult to interpret, but that values of less than .9 usually mean that the model can be improved substantially. Much of the value of these indices is based on the assumption that their behavior is independent of sample size.

One form of the incremental index, called type-1 incremental indices for present purposes, can be used to derive incremental fit indices from each of the stand alone indices described earlier. It has the form[c] of (t-n)/t for stand alone goodness of fit indices and (n-t)/n for stand alone badness of fit indices, where t is the value of a stand-alone index for the target model, and n is the value for the null model. For present purposes, incremental type-1 indices were defined in relation to the FF and X^2 (see Marsh, Balla & McDonald, 1988, for other type-1 indices), and are denoted by appending an I1 to each stand-alone index. The X^2I2 is more commonly known as the Bentler-Bonett Index (BBI) and Bollen (1986) described an index related to the FFI1 (see Marsh, Balla & McDonald, 1988). These are defined as:

(17)[c] $FFI1 = (FF_n - FF_t)/(FF_n)$.

(18) $X^2I1 = BBI = (X_n^2 - X_t^2)/(X_n^2)$.

IV. Relative, Incremental Type-2 Indices.

A second general form of the incremental fit indices described by Marsh, Balla, and McDonald (1988) is the type-2 incremental index. It has the form of (t-n)/(e - n) for stand alone goodness of fit indices and (n-t)/(n-e) for stand alone badness of fit indices, where t is the value of a stand-alone index for the target model, n is the value for the null model, and e is the expected value of the stand alone index if the target model is true. This second form of incremental index requires the expected value for a true model in addition to empirical values for the target and null models. In general, expected values for the stand-alone indices are not known for finite samples but can be estimated based on the asymptotic behavior of the indices. For example, many of the stand alone indices can be specified in terms of X^2 and the asymptotic expected value for the X^2 equals the df for the model. For purposes of the present investigation, incremental type-2 indices were derived from only the X^2/df and Dk stand-alone indices. These are denoted by appending an I2 to each of the stand-alone indices though the X^2/dfI2 is better known as the Tucker Lewis Index (Tucker & Lewis, 1973) and McDonald and Marsh (1988) refer to the DKI2 as the relative noncentrality index (RNI). These are defined as:

(19) $X^2/dfI2 = TLI = (X_n^2/df_n - X_t^2/df_t)/(X^2_n/df_n - [1.0]). = 1 - [\,(Dk_t/Dk_n) / (df_t/df_n)\,]$

(20)[b] $DkI2 = RNI = (Dk_n - Dk_t) / (Dk_n - 0)$

[The derivation of the TLI in terms of Dk is presented by Marsh and McDonald (1990). Note also that because the expected value of Dk for a true model is 0, the incremental type-1 and type-2 forms of this index are the same.]

IV. Parsimony Indices.

James et al. (1982; Mulaik, et al., 1989) also described an set of parsimony indices that invokes a penalty function for using additional parameters by multiplying a goodness of fit index by the ratio of the dfs for the null and target models: parsimony index = (df_t/df_n) x goodness of fit index. For present purposes I represent parsimony indices by placing a P in front of the fit index used in defining it. Using this general formulation, James et al. recommended a parsimony index based on the BBI defined as:

(21) $PBBI = (df_t/df_n) \times (X_n^2 - X_t^2)/(X_n^2)$.

Though not previously developed, Mulaik (personal communication) recommended that the parsimony index should more appropriately be constructed from the RNI instead of the BBI, and so I have also included this new parsimony index.

(22)[d] $PRNI = (df_t/df_n) \times (RNI)$

Additional parsimony indices could be derived for other goodness of fit indices, though this might not make sense for indices that already impose a penalty function (e.g., the AIC*, SC*, AIC, and SC).

---

[a] Steiger (1989) used a point estimate of the population noncentrality index that was obtained by an iterative procedure in his commercially available EzPATH program instead of Dk as used here. Steiger (p. 81) further noted, however, that one "could just as easily use the unbiased estimate. It is not yet clear which would be superior, but the differences bewteen the two approaches are likely to be trivial." Indices called GFI* and AGFI* here were referred to as $\tilde{}_1$ and $\tilde{}_2$ by Steiger (1989), though he also noted that these could be thought of as estimates of the population values of GFI and AGFI. [b] McDonald first described Dk and proposed Mc, two indices based on the noncentrality parameter, in late 1986 (see McDonald, 1990). Shortly after their development, in February of 1987, the incremental type-1 and type-2 forms of both these indices were proposed for purposes of the present investigation (also see McDonald & Marsh, 1990). Only the results of the DkI2 are actually presented here. DkI1 and DKI2 are mathematically identical whereas the pattern of empirical results based on the MCI2 were nearly identical to those based on DKI2. MCI1, because it was significantly related to sample size, was not pursued. Subsequently, in October 1988, McDonald and Marsh evaluated the mathematical properties of DKI2 more fully in research described in McDonald and Marsh (1990). For purposes of that paper and the present investigation, the index is referred to as the relative noncentrality index (RNI). In independent research, Bentler (1990) also developed incremental fit indices based on the noncentrality parameter that are apparently equivalent to those described here. Thus, credit for the development of what is called the RNI in the present investigation should appropriately be shared with Bentler (1990). [c] Marsh, Balla and McDonald (1988) used a slightly different notation to distinguish between the forms of incremental indices (type-I and type-II) based on goodness of fit indices and badness of fit indices. Although intended to be equivalent to the one presented here, it led to some confusion (e.g., Bollen, 1990, p. 256, footnote 1) so that Bollen inferred a slightly different formulation to the FFI1 (17) index presented here. [d] The parsimony index based on the RNI (PRNI) that is apparently first presented here was suggested by Mulaik (personal communication) based on his reading of McDonald and Marsh (1990) and Bentler (1990).

Appendix 2A

Mean Values for Goodness of Fit Indices For Each Model (M1-M8) At Different Sample Sizes: Simple Simulated Data

Model N	FF	LHR	χ^2	χ^2/df	RMR	RMS	GFI	AGFI	GFI*	AGFI*	AIC*	SC*	Ck	CN	Dk	MC	FFI1	BBI	TLI	NNI	PBBI	PNNI
M1 (df=27, PR=.75)																						
50	1.210	.551	59.369	2.196	.247	.154	.769	.615	.873	.788	1.930	2.618	2.133	38.199	.670	.722	.421	.421	.364	.519	.316	.389
100	.960	.621	95.078	3.521	.224	.158	.812	.687	.868	.780	1.320	1.789	1.364	47.283	.690	.711	.353	.353	.182	.386	.265	.289
200	.743	.692	147.869	5.476	.195	.148	.845	.741	.882	.803	.923	1.219	.933	62.703	.808	.741	.423	.423	.271	.453	.317	.340
400	.686	.710	273.978	10.147	.192	.150	.850	.751	.879	.799	.776	.956	.779	64.716	.619	.734	.420	.420	.245	.434	.315	.325
800	.722	.697	577.144	21.375	.207	.159	.845	.741	.867	.779	.767	.872	.768	61.088	.688	.709	.370	.370	.167	.375	.277	.281
1600	.644	.724	1030.357	38.161	.192	.152	.859	.766	.877	.796	.666	.727	.667	67.673	.627	.731	.426	.426	.240	.430	.319	.323
M2 (df=19, PR=.53)																						
50	.558	.758	27.341	1.439	.128	.088	.885	.729	.962	.911	1.598	2.592	1.891	62.608	.178	.916	.735	.735	.754	.865	.382	.457
100	.481	.786	47.713	2.511	.143	.122	.895	.751	.939	.856	1.001	1.679	1.066	70.423	.291	.865	.672	.672	.502	.736	.355	.388
200	.372	.831	74.156	3.902	.128	.117	.916	.802	.942	.864	.632	1.061	.647	100.545	.277	.872	.710	.710	.530	.751	.375	.396
400	.355	.837	141.949	7.471	.132	.126	.916	.800	.936	.848	.485	.745	.489	94.143	.308	.857	.699	.699	.467	.718	.369	.379
800	.357	.836	285.380	15.020	.141	.132	.916	.803	.931	.836	.422	.574	.423	92.332	.333	.846	.687	.687	.425	.696	.363	.367
1600	.346	.841	553.759	29.145	.140	.132	.919	.808	.930	.836	.378	.466	.379	94.519	.334	.846	.691	.691	.424	.696	.365	.367
M3 (df=12, PR=.33)																						
50	.234	.890	11.500	.958	.071	.036	.947	.8041.001		1.006	1.554	2.816	1.927	127.654	-.005	1.004	.883	.883	1.026	1.004	.294	.334
100	.123	.940	12.254	1.021	.054	.028	.971	.892	.999	.997	.783	1.643	.865	226.329	.003	.998	.913	.913	.978	.991	.304	.330
200	.056	.972	11.325	.943	.035	.016	.987	.9521.000		1.002	.386	.931	.406	462.976	-.003	1.001	.955	.955	1.015	1.004	.318	.334
400	.031	.984	12.452	1.037	.025	.014	.993	.974	.999	.999	.196	.525	.200	886.792	.001	.999	.973	.973	.997	.999	.324	.333
800	.012	.993	10.075	.839	.016	.005	.997	.9891.000		1.002	.095	.288	.096	2206.290	-.002	1.001	.989	.989	1.006	1.002	.329	.334
1600	.007	.996	12.151	1.012	.012	.006	.998	.9931.000		.999	.048	.159	.049	3501.754	.000	1.000	.993	.993	.999	.999	.331	.333
M4 (df=21, PR=.58)																						
50	.438	.804	21.472	1.022	.125	.037	.897	.779	.996	.992	1.398	2.316	1.669	86.738	.018	.992	.780	.780	.981	.982	.455	.573
100	.216	.898	21.419	1.020	.090	.027	.945	.883	.998	.997	.696	1.321	.755	201.366	.006	.997	.850	.850	.985	.989	.496	.577
200	.103	.949	20.630	.982	.063	.014	.971	.9391.000		1.000	.343	.739	.357	358.326	-.001	1.000	.918	.918	1.008	1.004	.535	.585
400	.055	.972	22.082	1.051	.045	.014	.985	.968	.999	.998	.175	.414	.178	691.496	.002	.998	.953	.953	.996	.997	.556	.582
800	.026	.986	21.084	1.004	.032	.006	.992	.9841.000		.999	.086	.226	.087	1479.291	.000	.999	.977	.977	1.000	1.000	.569	.583
1600	.013	.993	21.878	1.041	.022	.007	.996	.991	.999	.999	.043	.124	.043	3017.617	.000	.999	.987	.987	.999	.999	.576	.583
M5 (df=21, PR=.58)																						
50	.432	.807	21.211	1.010	.120	.037	.900	.787	.998	.995	1.392	2.310	1.663	90.477	.012	.996	.788	.788	1.012	1.000	.460	.583
100	.200	.905	19.835	.944	.082	.022	.950	.8931.002		1.005	.680	1.305	.739	212.464	-.009	1.005	.861	.861	1.009	1.003	.502	.585
200	.099	.951	19.865	.946	.059	.013	.973	.9431.001		1.002	.339	.735	.353	384.319	-.005	1.002	.921	.921	1.013	1.007	.537	.587
400	.053	.973	21.244	1.011	.043	.011	.986	.970	.999	.999	.173	.412	.176	717.051	.000	.999	.955	.955	.999	.999	.557	.583
800	.025	.987	20.471	.974	.032	.005	.992	.9841.000		1.000	.085	.226	.086	1524.725	-.000	1.000	.977	.977	1.001	1.000	.570	.583
1600	.012	.993	20.025	.953	.020	.005	.996	.9921.000		1.000	.042	.123	.042	3401.599	-.000	1.000	.988	.988	1.000	1.000	.576	.583
M6 (df=21, PR=.58)																						
50	.472	.791	23.158	1.102	.124	.051	.890	.764	.989	.977	1.432	2.350	1.703	84.407	.052	.976	.766	.766	.945	.961	.447	.560
100	.206	.902	20.454	.974	.083	.025	.949	.8911.001		1.002	.686	1.311	.745	211.743	-.003	1.002	.856	.856	.997	.996	.499	.581
200	.103	.949	20.608	.981	.061	.015	.972	.9401.000		1.000	.343	.739	.357	362.863	-.001	1.000	.917	.917	1.006	1.003	.535	.585
400	.052	.974	21.022	1.001	.043	.012	.986	.9701.000		.999	.172	.412	.176	730.708	.000	.999	.955	.955	1.000	.999	.557	.583
800	.025	.987	20.398	.971	.032	.006	.992	.9841.000		1.000	.085	.226	.086	1541.853	-.000	1.000	.977	.977	1.001	1.000	.570	.583
1600	.012	.993	20.623	.982	.021	.005	.996	.9921.000		1.000	.042	.123	.043	3144.322	-.000	1.000	.988	.988	1.000	1.000	.576	.583
M7 (df=18, PR=.50)																						
50	.355	.838	17.405	.966	.107	.036	.918	.7961.001		1.004	1.435	2.467	1.739	97.786	-.004	1.004	.822	.822	1.016	1.002	.411	.501
100	.170	.918	16.915	.939	.073	.021	.958	.8961.002		1.005	.710	1.414	.777	225.918	-.009	1.005	.881	.881	1.011	1.003	.440	.501
200	.087	.957	17.327	.962	.053	.014	.977	.9431.000		1.001	.357	.802	.372	390.080	-.002	1.001	.931	.931	1.010	1.005	.465	.502
400	.044	.977	18.446	1.024	.038	.013	.988	.970	.999	.999	.181	.450	.185	757.099	.001	.999	.961	.961	.998	.999	.480	.499
800	.022	.988	17.877	.993	.028	.005	.993	.9841.000		1.000	.089	.248	.090	1558.645	-.000	1.000	.980	.980	1.000	1.000	.490	.500
1600	.011	.994	18.508	1.028	.019	.006	.996	.992	.999	.999	.045	.136	.045	3264.666	.000	.999	.989	.989	.999	.999	.494	.499
M8 (df=24, PR=.67)																						
50	.521	.771	25.544	1.064	.137	.037	.878	.772	.991	.984	1.361	2.164	1.598	79.506	.041	.981	.742	.742	.970	.972	.495	.648
100	.245	.885	24.266	1.011	.096	.029	.938	.883	.999	.998	.665	1.212	.717	195.285	.005	.998	.830	.830	.988	.990	.553	.660
200	.117	.943	23.336	.972	.068	.016	.967	.9381.000		1.001	.327	.673	.339	357.706	-.002	1.001	.907	.907	1.010	1.006	.605	.670
400	.062	.969	24.792	1.033	.049	.013	.983	.968	.999	.999	.167	.376	.170	667.535	.002	.999	.947	.947	.997	.998	.631	.665
800	.028	.985	22.729	.947	.034	.005	.991	.9831.000		1.000	.080	.203	.081	1519.209	-.001	1.000	.975	.975	1.002	1.001	.650	.667
1600	.014	.992	23.380	.974	.024	.005	.995	.9921.000		1.000	.040	.111	.041	3114.055	-.000	1.000	.987	.987	1.000	1.000	.658	.666

Note: PR=Parsimony Ratio. See Appendix 1 for a description of the Goodness of Fit indices.

Appendix 2B

Mean Values for Goodness of Fit Indices For Each Model (M1-M8) At Different Sample Sizes: Complex Simulated Data

Model N	FF	LHR	χ^2	χ^2 /df	RMR	RMS	GFI	AGFI	GFI*	AGFI*	AIC*	BC*	Ck	CN	Dk	MC	FFI1	BBI	TLI	RNI	PBBI	PRNI	M1
(df=27, PR=.75)																							
50	1.139	.570	55.820	2.067	.220	.145	.780	.634	.884	.807	1.859	2.567	2.062	40.406	.599	.746	.501	.501	.498	.620	.376	.465	
100	.767	.683	75.999	2.814	.183	.133	.841	.736	.901	.835	1.127	1.596	1.172	59.291	.497	.782	.522	.522	.472	.603	.391	.483	
200	.663	.719	131.973	4.887	.173	.138	.856	.761	.895	.825	.843	1.148	.853	67.745	.528	.769	.551	.551	.458	.593	.413	.445	
400	.591	.744	236.055	8.743	.168	.138	.866	.777	.895	.826	.681	.861	.684	74.399	.524	.770	.565	.565	.451	.588	.424	.441	
800	.599	.741	476.588	17.725	.171	.144	.862	.770	.888	.814	.644	.749	.644	73.456	.565	.754	.576	.576	.449	.587	.432	.440	
1600	.583	.747	932.179	34.525	.170	.144	.865	.775	.888	.813	.605	.666	.605	74.820	.566	.753	.572	.572	.437	.577	.429	.433	
M2 (df=19, PR=.53)																							
50	.581	.750	28.480	1.498	.130	.094	.878	.710	.958	.902	1.621	2.615	1.914	64.529	.201	.908	.747	.747	.792	.885	.394	.467	
100	.386	.825	38.239	2.012	.119	.098	.912	.792	.958	.901	.906	1.583	.970	89.860	.196	.907	.757	.757	.782	.841	.399	.444	
200	.319	.853	63.518	3.343	.113	.107	.923	.817	.952	.888	.579	1.008	.594	108.699	.224	.894	.783	.783	.672	.827	.413	.436	
400	.306	.858	122.110	6.426	.117	.116	.924	.820	.945	.871	.436	.695	.439	109.333	.258	.879	.775	.775	.615	.797	.409	.420	
800	.279	.869	223.221	11.748	.113	.115	.929	.833	.946	.872	.344	.496	.345	119.787	.255	.880	.802	.802	.646	.813	.433	.429	
1600	.301	.860	482.393	25.389	.123	.123	.923	.819	.939	.856	.334	.421	.334	108.999	.289	.865	.778	.778	.589	.783	.410	.413	
M3 (df=12, PR=.33)																							
50	.222	.895	10.906	.908	.068	.030	.948	.808	1.004	1.015	1.542	2.804	1.914	119.396	-.017	1.009	.900	.900	1.055	1.014	.300	.338	
100	.126	.939	12.486	1.040	.055	.029	.971	.893	1.998	.995	.786	1.645	.867	198.563	.006	.997	.921	.921	.995	.997	.307	.332	
200	.060	.970	12.019	1.001	.035	.019	.986	.950	.999	.999	.390	.934	.409	443.535	.000	.999	.958	.958	1.000	1.000	.319	.333	
400	.031	.984	12.532	1.044	.026	.013	.993	.973	.999	.998	.196	.525	.201	811.572	.001	.999	.977	.977	.998	.999	.325	.333	
800	.016	.991	13.467	1.122	.019	.012	.996	.985	.999	.998	.099	.292	.100	1494.087	.001	.999	.987	.987	.995	.998	.329	.332	
1600	.007	.996	12.671	1.055	.013	.007	.998	.993	.999	.999	.049	.160	.049	3144.355	.000	.999	.994	.994	.999	.999	.331	.333	
M4 (df=21, PR=.58)																							
50	.541	.765	26.518	1.262	.149	.066	.874	.732	.975	.946	1.501	2.419	1.772	73.489	.121	.944	.764	.764	.903	.938	.445	.547	
100	.289	.865	28.673	1.365	.106	.056	.928	.847	.982	.963	.769	1.394	.829	126.264	.079	.961	.817	.817	.894	.936	.476	.546	
200	.163	.921	32.550	1.550	.082	.048	.956	.905	.987	.972	.403	.799	.417	225.180	.058	.971	.889	.889	.927	.957	.510	.558	
400	.108	.947	43.156	2.055	.071	.048	.968	.931	.987	.973	.228	.467	.231	361.166	.055	.972	.921	.921	.926	.956	.537	.558	
800	.074	.963	59.843	2.849	.060	.047	.977	.950	.989	.977	.134	.275	.135	492.186	.048	.976	.946	.946	.938	.964	.552	.562	
1600	.071	.965	113.878	5.422	.061	.052	.976	.949	.987	.972	.101	.181	.101	501.198	.058	.971	.947	.947	.925	.956	.552	.558	
M5 (df=21, PR=.58)																							
50	.520	.773	25.511	1.214	.134	.059	.879	.742	.979	.955	1.480	2.398	1.751	75.242	.100	.953	.771	.771	.917	.945	.450	.551	
100	.276	.871	27.382	1.303	.104	.046	.930	.851	.985	.969	.756	1.381	.815	136.012	.066	.968	.828	.828	.920	.952	.483	.555	
200	.177	.915	35.266	1.679	.086	.056	.950	.893	.984	.966	.417	.813	.431	207.619	.072	.964	.878	.878	.905	.944	.512	.551	
400	.108	.947	43.157	2.055	.069	.050	.969	.933	.987	.973	.228	.467	.231	341.292	.055	.972	.920	.920	.925	.956	.537	.558	
800	.081	.960	65.113	3.100	.064	.051	.975	.946	.987	.974	.141	.282	.142	436.778	.055	.972	.942	.942	.930	.959	.549	.559	
1600	.074	.963	119.030	5.668	.063	.053	.976	.949	.986	.971	.104	.185	.104	486.017	.061	.969	.945	.945	.921	.954	.551	.556	
M6 (df=21, PR=.58)																							
50	.519	.773	25.453	1.212	.142	.066	.881	.745	.979	.956	1.479	2.397	1.750	81.138	.099	.954	.777	.777	.944	.961	.453	.560	
100	.318	.853	31.561	1.502	.110	.065	.922	.834	.976	.949	.798	1.424	.858	116.020	.108	.947	.799	.799	.853	.913	.466	.532	
200	.188	.910	37.580	1.789	.087	.060	.951	.895	.981	.961	.428	.824	.442	195.864	.083	.959	.870	.870	.889	.935	.507	.545	
400	.119	.942	47.777	2.275	.072	.055	.967	.931	.985	.968	.239	.479	.243	312.890	.067	.967	.912	.912	.909	.947	.532	.552	
800	.095	.953	76.367	3.636	.066	.057	.973	.942	.984	.967	.155	.296	.156	377.155	.069	.966	.931	.931	.912	.949	.543	.553	
1600	.092	.954	148.289	7.061	.068	.061	.973	.943	.982	.962	.122	.203	.122	384.827	.079	.961	.932	.932	.898	.940	.543	.548	
M7 (df=18, PR=.50)																							
50	.383	.827	18.787	1.042	.110	.035	.907	.768	.995	.989	1.462	2.495	1.767	90.227	.022	.990	.830	.830	.996	.992	.415	.496	
100	.189	.909	18.765	1.042	.079	.025	.953	.884	.998	.994	.729	1.432	.796	173.138	.009	.995	.881	.881	.995	.996	.440	.498	
200	.100	.950	20.062	1.114	.058	.027	.973	.934	.997	.994	.370	.816	.386	349.820	.010	.994	.931	.931	.987	.993	.465	.496	
400	.046	.977	18.379	1.021	.038	.013	.988	.971	.999	.999	.181	.450	.184	740.417	.001	.999	.966	.966	.999	.999	.483	.499	
800	.024	.987	19.565	1.086	.027	.010	.993	.984	.999	.998	.092	.250	.092	1406.013	.002	.999	.982	.982	.996	.998	.491	.499	
1600	.011	.994	18.523	1.029	.019	.004	.996	.992	.999	.999	.045	.136	.045	2826.609	.000	.999	.991	.991	.999	.999	.495	.499	
M8 (df=24, PR=.67)																							
50	.455	.723	32.140	1.339	.170	.078	.848	.715	.964	.932	1.495	2.299	1.732	66.851	.175	.920	.716	.716	.875	.910	.477	.606	
100	.358	.836	35.524	1.480	.121	.066	.912	.836	.974	.952	.778	1.325	.830	113.817	.118	.942	.774	.774	.861	.906	.516	.604	
200	.226	.893	45.053	1.877	.098	.065	.939	.886	.977	.956	.436	.782	.448	178.012	.106	.948	.845	.845	.878	.918	.563	.612	
400	.155	.925	61.999	2.583	.086	.061	.956	.918	.979	.961	.260	.469	.263	268.222	.095	.953	.886	.886	.888	.928	.590	.616	
800	.105	.948	84.105	3.504	.072	.055	.968	.938	.983	.969	.157	.280	.158	383.674	.075	.963	.925	.925	.917	.944	.616	.629	
1600	.118	.942	189.808	7.908	.079	.065	.964	.933	.977	.957	.145	.215	.145	334.848	.103	.949	.912	.912	.884	.922	.608	.615	

Note: PR=Parsimony Ratio. See Appendix 1 for a description of the Goodness of Fit indices.